Weak semantic context helps phonetic learning in a model of infant language acquisition

Stella Frank
sfrank@inf.ed.ac.uk
ILCC, School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, UK

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

Sharon Goldwater
sgwater@inf.ed.ac.uk
ILCC, School of Informatics
University of Edinburgh
Edinburgh, EH8 9AB, UK

Abstract

Learning phonetic categories is one of the first steps to learning a language, yet is hard to do using only distributional phonetic information. Semantics could potentially be useful, since words with different meanings have distinct phonetics, but it is unclear how many word meanings are known to infants learning phonetic categories. We show that attending to a weaker source of semantics, in the form of a distribution over topics in the current context, can lead to improvements in phonetic category learning. In our model, an extension of a previous model of joint word-form and phonetic category inference, the probability of word-forms is topic-dependent, enabling the model to find significantly better phonetic vowel categories and word-forms than a model with no semantic knowledge.

1 Introduction

Infants begin learning the phonetic categories of their native language in their first year (Kuhl et al., 1992; Polka and Werker, 1994; Werker and Tees, 1984). In theory, semantic information could offer a valuable cue for phoneme induction[1] by helping infants distinguish between minimal pairs, as linguists do (Trubetzkoy, 1939). However, due to a widespread assumption that infants do not know the meanings of many words at the age when they are learning phonetic categories (see Swingley, 2009 for a review), most recent models of early phonetic category acquisition have explored the phonetic learning problem in the absence of semantic information (de Boer and Kuhl, 2003; Dillon et al., 2013; Feldman et al., 2013a; McMurray et al., 2009; Vallabha et al., 2007).

Models without any semantic information are likely to underestimate infants' ability to learn phonetic categories. Infants learn language in the wild, and quickly attune to the fact that words have (possibly unknown) meanings. The extent of infants' semantic knowledge is not yet known, but existing evidence shows that six-month-olds can associate some words with their referents (Bergelson and Swingley, 2012; Tincoff and Jusczyk, 1999, 2012), leverage non-acoustic contexts such as objects or articulations to distinguish similar sounds (Teinonen et al., 2008; Yeung and Werker, 2009), and map meaning (in the form of objects or images) to new word-forms in some laboratory settings (Friedrich and Friederici, 2011; Gogate and Bahrick, 2001; Shukla et al., 2011). These findings indicate that young infants are sensitive to co-occurrences between linguistic stimuli and at least some aspects of the world.

In this paper we explore the potential contribution of semantic information to phonetic learning by formalizing a model in which learners attend to the word-level context in which phones appear (as in the lexical-phonetic learning model of Feldman et al., 2013a) and also to the situations in which word-forms are used. The modeled situations consist of combinations of categories of salient activities or objects, similar to the activity contexts explored by Roy et al. (2012), e.g.,'getting dressed' or 'eating breakfast'. We assume that child learners are able to infer a representation of the situational context from their non-linguistic environment. However, in our simulations we approximate the environmental information by running a topic model (Blei et al., 2003) over a corpus of child-directed speech to infer a topic distribution for each situation. These topic distributions are then used as input to our model to represent situational contexts.

The situational information in our model is simi-

[1] The models in this paper do not distinguish between phonetic and phonemic categories, since they do not capture phonological processes (and there are also none present in our synthetic data). We thus use the terms interchangeably.

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 1073–1083,
Baltimore, Maryland, USA, June 23-25 2014.

lar to that assumed by theories of cross-situational word learning (Frank et al., 2009; Smith and Yu, 2008; Yu and Smith, 2007), but our model does not require learners to map individual words to their referents. Even in the absence of word-meaning mappings, situational information is potentially useful because similar-sounding words uttered in similar situations are more likely to be tokens of the same lexeme (containing the same phones) than similar-sounding words uttered in different situations.

In simulations of vowel learning, inspired by Vallabha et al. (2007) and Feldman et al. (2013a), we show a clear improvement over previous models in both phonetic and lexical (word-form) categorization when situational context is used as an additional source of information. This improvement is especially noticeable when the word-level context is providing less information, arguably the more realistic setting. These results demonstrate that relying on situational co-occurrence can improve phonetic learning, even if learners do not yet know the meanings of individual words.

2 Background and overview of models

Infants attend to distributional characteristics of their input (Maye et al., 2002, 2008), leading to the hypothesis that phonetic categories could be acquired on the basis of bottom-up distributional learning alone (de Boer and Kuhl, 2003; Vallabha et al., 2007; McMurray et al., 2009). However, this would require sound categories to be well separated, which often is not the case—for example, see Figure 1, which shows the English vowel space that is the focus of this paper.

Recent work has investigated whether infants could overcome such distributional ambiguity by incorporating top-down information, in particular, the fact that phones appear within words. At six months, infants begin to recognize word-forms such as their name and other frequently occurring words (Mandel et al., 1995; Jusczyk and Hohne, 1997), without necessarily linking a meaning to these forms. This "protolexicon" can help differentiate phonetic categories by adding word contexts in which certain sound categories appear (Swingley, 2009; Feldman et al., 2013b). To explore this idea further, Feldman et al. (2013a) implemented the Lexical-Distributional (LD) model, which jointly learns a set of phonetic vowel categories and a set of word-forms containing those categories. Simulations showed that the use of lexical context greatly improved phonetic learning.

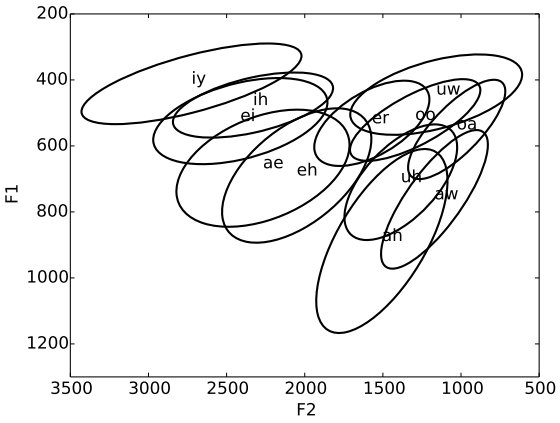

Figure 1: The English vowel space (generated from Hillenbrand et al. (1995), see Section 6.2), plotted using the first two formants.

Our own Topic-Lexical-Distributional (TLD) model extends the LD model to include an additional type of context: the situations in which words appear. To motivate this extension and clarify the differences between the models, we now provide a high-level overview of both models; details are given in Sections 3 and 4.

2.1 Overview of LD model

Both the LD and TLD models are computational-level models of phonetic (specifically, vowel) categorization where phones (vowels) are presented to the model in the context of words.[2] The task is to infer a set of phonetic categories and a set of lexical items on the basis of the data observed for each word token x_i. In the original LD model, the observations for token x_i are its frame f_i, which consists of a list of consonants and slots for vowels, and the list of vowel tokens $\boldsymbol{w}_i$. (The TLD model includes additional observations, described below.) A single vowel token, w_{ij}, is a two dimensional vector representing the first two formants (peaks in the frequency spectrum, ordered from lowest to highest). For example, a token of the word *kitty* would have the frame $f_i = \texttt{k\_t\_}$, containing two consonant phones, /k/ and /t/, with two vowel phone slots in between, and two vowel formant vectors,

[2]For a related model that also tackles the word segmentation problem, see Elsner et al. (2013). In a model of phonological learning, Fourtassi and Dupoux (submitted) show that semantic context information similar to that used here remains useful despite segmentation errors.

$w_{i0} = [464, 2294]$ and $w_{i1} = [412, 2760]$.[3]

Given the data, the model must assign each vowel token to a vowel category, $w_{ij} = c$. Both the LD and the TLD models do this using intermediate lexemes, $\boldsymbol{\ell}$, which contain vowel category assignments, $v_{\ell j} = c$, as well as a frame f_ℓ. If a word token is assigned to a lexeme, $x_i = \ell$, the vowels within the word are assigned to that lexeme's vowel categories, $w_{ij} = v_{\ell j} = c$.[4] The word and lexeme frames must match, $f_i = f_\ell$.

Lexical information helps with phonetic categorization because it can disambiguate highly overlapping categories, such as the *ae* and *eh* categories in Figure 1. A purely distributional learner who observes a cluster of data points in the *ae-eh* region is likely to assume all these points belong to a single category because the distributions of the categories are so similar. However, a learner who attends to lexical context will notice a difference: contexts that only occur with *ae* will be observed in one part of the *ae-eh* region, while contexts that only occur with *eh* will be observed in a different (though partially overlapping) space. The learner then has evidence of two different categories occurring in different sets of lexemes.

Simulations with the LD model show that using lexical information to constrain phonetic learning can greatly improve categorization accuracy (Feldman et al., 2013a), but it can also introduce errors. When two word tokens contain the same consonant frame but different vowels (i.e., minimal pairs), the model is more likely to categorize those two vowels together. Thus, the model has trouble distinguishing minimal pairs. Although young children also have trouble with minimal pairs (Stager and Werker, 1997; Thiessen, 2007), the LD model may overestimate the degree of the problem. We hypothesize that if a learner is able to associate words with the contexts of their use (as children likely are), this could provide a weak source of information for disambiguating minimal pairs even without knowing their exact meanings. That is, if the learner hears kV_1t and kV_2t in different situational contexts, they are likely to be different lexical items (and V_1 and V_2 different phones), despite the lexical similarity between them.

[3] In simulations we also experiment with frames in which consonants are not represented perfectly.

[4] The notation is overloaded: w_{ij} refers both to the vowel formants and the vowel category assignments, and x_i refers to both the token identity and its assignment to a lexeme.

2.2 Overview of TLD model

To demonstrate the benefit of situational information, we develop the Topic-Lexical-Distributional (TLD) model, which extends the LD model by assuming that words appear in *situations* analogous to documents in a topic model. Each situation h is associated with a mixture of topics θ_h, which is assumed to be observed. Thus, for the ith token in situation h, denoted x_{hi}, the observed data will be its frame f_{hi}, vowels $\boldsymbol{w}_{hi}$, and topic vector θ_h.

From an acquisition perspective, the observed topic distribution represents the child's knowledge of the context of the interaction: she can distinguish bathtime from dinnertime, and is able to recognize that some topics appear in certain contexts (e.g. animals on walks, vegetables at dinnertime) and not in others (few vegetables appear at bathtime). We assume that the child would learn these topics from observing the world around her and the co-occurrences of entities and activities in the world. Within any given situation, there might be a mixture of different (actual or possible) topics that are salient to the child. We assume further that as the child learns the language, she will begin to associate specific words with each topic as well.

Thus, in the TLD model, the words used in a situation are topic-dependent, implying meaning, but without pinpointing specific referents. Although the model observes the distribution of topics in each situation (corresponding to the child observing her non-linguistic environment), it must learn to associate each (phonetically and lexically ambiguous) word token with a particular topic from that distribution. The occurrence of similar-sounding words in different situations with mostly non-overlapping topics will provide evidence that those words belong to different topics and that they are therefore different lexemes. Conversely, potential minimal pairs that occur in situations with similar topic distributions are more likely to belong to the same topic and thus the same lexeme.

Although we assume that children infer topic distributions from the non-linguistic environment, we will use transcripts from CHILDES to create the word/phone learning input for our model. These transcripts are not annotated with environmental context, but Roy et al. (2012) found that topics learned from similar transcript data using a topic model were strongly correlated with immediate activities and contexts. We therefore obtain the topic distributions used as input to the TLD model by

training an LDA topic model (Blei et al., 2003) on a superset of the child-directed transcript data we use for lexical-phonetic learning, dividing the transcripts into small sections (the 'documents' in LDA) that serve as our distinct situations $\boldsymbol{h}$. As noted above, the learned document-topic distributions $\boldsymbol{\theta}$ are treated as observed variables in the TLD model to represent the situational context. The topic-word distributions learned by LDA are discarded, since these are based on the (correct and unambiguous) words in the transcript, whereas the TLD model is presented with phonetically ambiguous versions of these word tokens and must learn to disambiguate them and associate them with topics.

3 Lexical-Distributional Model

In this section we describe more formally the generative process for the LD model (Feldman et al., 2013a), a joint Bayesian model over phonetic categories and a lexicon, before describing the TLD extension in the following section.

The set of phonetic categories and the lexicon are both modeled using non-parametric Dirichlet Process priors, which return a potentially infinite number of categories or lexemes. A DP is parametrized as $DP(\alpha, H)$, where α is a real-valued hyperparameter and H is a base distribution. H may be continuous, as when it generates phonetic categories in formant space, or discrete, as when it generates lexemes as a list of phonetic categories.

A draw from a DP, $G \sim DP(\alpha, H)$, returns a distribution over a set of draws from H, i.e., a discrete distribution over a set of categories or lexemes generated by H. In the mixture model setting, the category assignments are then generated from G, with the datapoints themselves generated by the corresponding components from H. If H is infinite, the support of the DP is likewise infinite. During inference, we marginalize over G.

3.1 Phonetic Categories: IGMM

Following previous models of vowel learning (de Boer and Kuhl, 2003; Vallabha et al., 2007; McMurray et al., 2009; Dillon et al., 2013) we assume that vowel tokens are drawn from a Gaussian mixture model. The Infinite Gaussian Mixture Model (IGMM) (Rasmussen, 2000) includes a DP prior, as described above, in which the base distribution H_C generates multivariate Gaussians drawn from a Normal Inverse-Wishart prior.[5] Each observation, a formant vector w_{ij}, is drawn from the Gaussian corresponding to its category assignment c_{ij}:

$$\mu_c, \Sigma_c \sim H_C = \mathit{NIW}(\mu_0, \Sigma_0, \nu_0) \quad (1)$$

$$G_C \sim DP(\alpha_c, H_C) \quad (2)$$

$$c_{ij} \sim G_C \quad (3)$$

$$w_{ij} | c_{ij} = c \sim N(\mu_c, \Sigma_c) \quad (4)$$

The above model generates a category assignment c_{ij} for each vowel token w_{ij}. This is the baseline IGMM model, which clusters vowel tokens using bottom-up distributional information only; the LD model adds top-down information by assigning categories in the lexicon, rather than on the token level.

3.2 Lexicon

In the LD model, vowel phones appear within words drawn from the lexicon. Each such lexeme is represented as a frame plus a list of vowel categories $\boldsymbol{v}_\ell$. Lexeme assignments for each token are drawn from a DP with a lexicon-generating base distribution H_L. The category for each vowel token in the word is determined by the lexeme; the formant values are drawn from the corresponding Gaussian as in the IGMM:

$$G_L \sim DP(\alpha_l, H_L) \quad (5)$$

$$x_i = \ell \sim G_L \quad (6)$$

$$w_{ij} | v_{\ell j} = c \sim N(\mu_c, \Sigma_c) \quad (7)$$

H_L generates lexemes by first drawing the number of phones from a geometric distribution and the number of consonant phones from a binomial distribution. The consonants are then generated from a DP with a uniform base distribution (but note they are fixed at inference time, i.e., are observed categorically), while the vowel phones $\boldsymbol{v}_\ell$ are generated by the IGMM DP above, $v_{\ell j} \sim G_C$.

Note that two draws from H_L may result in identical lexemes; these are nonetheless considered to be separate (homophone) lexemes.

4 Topic-Lexical-Distributional Model

The TLD model retains the IGMM vowel phone component, but extends the lexicon of the LD model by adding topic-specific lexicons, which capture the notion that lexeme probabilities are topic-dependent. Specifically, the TLD model replaces

[5] This compound distribution is equivalent to $\Sigma_c \sim \mathit{IW}(\Sigma_0, \nu_0)$, $\mu_c | \Sigma_c \sim N(\mu_0, \frac{\Sigma_c}{\nu_0})$

the Dirichlet Process lexicon with a Hierarchical Dirichlet Process (HDP; Teh (2006)). In the HDP lexicon, a top-level global lexicon is generated as in the LD model. Topic-specific lexicons are then drawn from the global lexicon, containing a subset of the global lexicon (but since the size of the global lexicon is unbounded, so are the topic-specific lexicons). These topic-specific lexicons are used to generate the tokens in a similar manner to the LD model. There are a fixed number of lower level topic-lexicons; these are matched to the number of topics in the LDA model used to infer the topic distributions (see Section 6.4).

More formally, the global lexicon is generated as a top-level DP: $G_L \sim DP(\alpha_l, H_L)$ (see Section 3.2; remember H_L includes draws from the IGMM over vowel categories). G_L is in turn used as the base distribution in the topic-level DPs, $G_k \sim DP(\alpha_k, G_L)$. In the Chinese Restaurant Franchise metaphor often used to describe HDPs, G_L is a global menu of dishes (lexemes). The topic-specific lexicons are restaurants, each with its own distribution over dishes; this distribution is defined by seating customers (word tokens) at *tables*, each of which serves a single dish from the menu: all tokens x at the same table t are assigned to the same lexeme ℓ_t. Inference (Section 5) is defined in terms of tables rather than lexemes; if multiple tables draw the same dish from G_L, tokens at these tables share a lexeme.

In the TLD model, tokens appear within situations, each of which has a distribution over topics θ_h. Each token x_{hi} has a co-indexed topic assignment variable, z_{hi}, drawn from θ_h, designating the topic-lexicon from which the table for x_{hi} is to be drawn. The formant values for w_{hij} are drawn in the same way as in the LD model, given the lexeme assignment at x_{hi}. This results in the following model, shown in Figure 2:

$$G_L \sim DP(\alpha_l, H_L) \quad (8)$$

$$G_k \sim DP(\alpha_k, G_L) \quad (9)$$

$$z_{hi} \sim Mult(\theta_h) \quad (10)$$

$$x_{hi} = t | z_{hi} = k \sim G_k \quad (11)$$

$$w_{hij} | x_{hi} = t, v_{\ell_t j} = c \sim N(\mu_c, \Sigma_c) \quad (12)$$

5 Inference: Gibbs Sampling

We use Gibbs sampling to infer three sets of variables in the TLD model: assignments to vowel categories in the lexemes, assignments of tokens to topics, and assignments of tokens to tables (from which the assignment to lexemes can be read off).

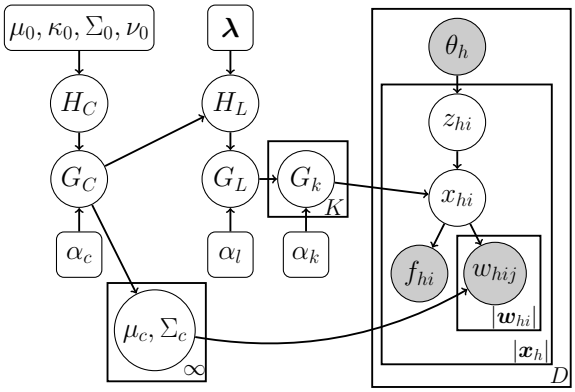

Figure 2: TLD model, depicting, from left to right, the IGMM component, the LD lexicon component, the topic-specific lexicons, and finally the token x_{hi}, appearing in document h, with observed vowel formants w_{hij} and frame f_{hi}. The lexeme assignment x_{hi} and the topic assignment z_{hi} are inferred, the latter using the observed document-topic distribution θ_h. Note that f_i is deterministic given the lexeme assignment. Squared nodes depict hyperparameters. $\boldsymbol{\lambda}$ is the set of hyperparameters used by H_L when generating lexical items (see Section 3.2).

5.1 Sampling lexeme vowel categories

Each vowel in the lexicon must be assigned to a category in the IGMM. The posterior probability of a category assignment is composed of the DP prior over categories and the likelihood of the observed vowels belonging to that category. We use $\boldsymbol{w}_{\ell j}$ to denote the set of vowel formants at position j in words that have been assigned to lexeme ℓ. Then,

$$P(v_{\ell j} = c | \boldsymbol{w}, \boldsymbol{x}, \boldsymbol{\ell}^{\setminus \ell}) \propto P(v_{\ell j} = c | \boldsymbol{\ell}^{\setminus \ell}) p(\boldsymbol{w}_{\ell j} | v_{\ell j} = c, \boldsymbol{w}^{\setminus \ell j}) \quad (13)$$

The first (DP prior) factor is defined as:

$$P(v_{\ell j} = c | \boldsymbol{v}^{\setminus \ell j}) = \begin{cases} \frac{n_c}{\sum_c n_c + \alpha_c} & \text{if } c \text{ exists} \\ \frac{\alpha_c}{\sum_c n_c + \alpha_c} & \text{if } c \text{ new} \end{cases} \quad (14)$$

where n_c is the number of other vowels in the lexicon, $\boldsymbol{v}^{\setminus lj}$, assigned to category c. Note that there is always positive probability of creating a new category.

The likelihood of the vowels is calculated by marginalizing over all possible means and variances of the Gaussian category parameters, given

the NIW prior. For a single point (if $|\boldsymbol{w}_{\ell j}| = 1$), this predictive posterior is in the form of a Student-t distribution; for the more general case see Feldman et al. (2013a), Eq. B3.

5.2 Sampling table & topic assignments

We jointly sample $\boldsymbol{x}$ and $\boldsymbol{z}$, the variables assigning tokens to tables and topics. Resampling the table assignment includes the possibility of changing to a table with a different lexeme or drawing a new table with a previously seen or novel lexeme. The joint conditional probability of a table and topic assignment, given all other current token assignments, is:

$$\begin{aligned} &P(x_{hi} = t, z_{hi} = k | \boldsymbol{w}_{hi}, \theta_h, \boldsymbol{t}^{\setminus hi}, \boldsymbol{\ell}, \boldsymbol{w}^{\setminus hi}) \\ &\quad = P(k|\theta_h) P(t|k, \ell_t, \boldsymbol{t}^{\setminus hi}) \\ &\qquad \prod_{c \in C} p(\boldsymbol{w}_{hi\cdot} | v_{\ell_t \cdot} = c, \boldsymbol{w}^{\setminus hi}) \end{aligned} \tag{15}$$

The first factor, the prior probability of topic k in document h, is given by θ_{hk} obtained from the LDA. The second factor is the prior probability of assigning word x_i to table t with lexeme ℓ given topic k. It is given by the HDP, and depends on whether the table t exists in the HDP topic-lexicon for k and, likewise, whether any table in the topic-lexicon has the lexeme ℓ:

$$P(t|k, \ell, \boldsymbol{t}^{\setminus hi}) \propto \begin{cases} \frac{n_{kt}}{n_k + \alpha_k} & \text{if } t \text{ in } k \\ \frac{\alpha_k}{n_k + \alpha_k} \frac{m_\ell}{m + \alpha_l} & \text{if } t \text{ new, } \ell \text{ known} \\ \frac{\alpha_k}{n_k + \alpha_k} \frac{\alpha_\ell}{m + \alpha_l} & \text{if } t \text{ and } \ell \text{ new} \end{cases} \tag{16}$$

Here n_{kt} is the number of other tokens at table t, n_k are the total number of tokens in topic k, m_ℓ is the number of tables across all topics with the lexeme ℓ, and m is the total number of tables.

The third factor, the likelihood of the vowel formants $\boldsymbol{w}_{hi}$ in the categories given by the lexeme $\boldsymbol{v}_l$, is of the same form as the likelihood of vowel categories when resampling lexeme vowel assignments. However, here it is calculated over the set of vowels in the token assigned to each vowel category (i.e., the vowels at indices where $v_{\ell_t \cdot} = c$). For a new lexeme, we approximate the likelihood using 100 samples drawn from the prior, each weighted by $\alpha/100$ (Neal, 2000).

5.3 Hyperparameters

The three hyperparameters governing the HDP over the lexicon, α_l and α_k, and the DP over vowel categories, α_c, are estimated using a slice sampler. The remaining hyperparameters for the vowel category and lexeme priors are set to the same values used by Feldman et al. (2013a).

6 Experiments

6.1 Corpus

We test our model on situated child directed speech, taken from the C1 section of the Brent corpus in CHILDES (Brent and Siskind, 2001; MacWhinney, 2000). This corpus consists of transcripts of speech directed at infants between the ages of 9 and 15 months, captured in a naturalistic setting as parent and child went about their day. This ensures variability of situations.

Utterances with unintelligible words or quotes are removed. We restrict the corpus to content words by retaining only words tagged as `adj`, `n`, `part` and `v` (adjectives, nouns, particles, and verbs). This is in line with evidence that infants distinguish content and function words on the basis of acoustic signals (Shi and Werker, 2003). Vowel categorization improves when attending only to more prosodically and phonologically salient tokens (Adriaans and Swingley, 2012), which generally appear within content, not function words. The final corpus consists of 13138 tokens and 1497 word types.

6.2 Hillenbrand Vowels

The transcripts do not include phonetic information, so, following Feldman et al. (2013a), we synthesize the formant values using data from Hillenbrand et al. (1995). This dataset consists of a set of 1669 manually gathered formant values from 139 American English speakers (men, women and children) for 12 vowels. For each vowel category, we construct a Gaussian from the mean and covariance of the datapoints belonging to that category, using the first and second formant values measured at steady state. We also construct a second dataset using only datapoints from adult female speakers.

Each word in the dataset is converted to a phonemic representation using the CMU pronunciation dictionary, which returns a sequence of Arpabet phoneme symbols. If there are multiple possible pronunciations, the first one is used. Each vowel phoneme in the word is then replaced by formant values drawn from the corresponding Hillenbrand Gaussian for that vowel.

6.3 Merging Consonant Categories

The Arpabet encoding used in the phonemic representation includes 24 consonants. We construct datasets both using the full set of consonants—the 'C24' dataset—and with less fine-grained consonant categories. Distinguishing all consonant categories assumes perfect learning of consonants prior to vowel categorization and is thus somewhat unrealistic (Polka and Werker, 1994), but provides an upper limit on the information that word-contexts can give.

In the 'C15' dataset, the voicing distinction is collapsed, leaving 15 consonant categories. The collapsed categories are B/P, G/K, D/T, CH/JH, V/F, TH/DH, S/Z, SH/ZH, R/L while HH, M, NG, N, W, Y remain separate phonemes. This dataset mirrors the finding in Mani and Plunkett (2010) that 12 month old infants are not sensitive to voicing mispronunciations.

The 'C6' dataset distinguishes between only 6 coarse consonant phonemes, corresponding to stops (B,P,G,K,D,T), affricates (CH,JH), fricatives (V, F, TH, DH, S, Z, SH, ZH, HH), nasals (M, NG, N), liquids (R, L), and semivowels/glides (W, Y). This dataset makes minimal assumptions about the category categories that infants could use in this learning setting.

Decreasing the number of consonants increases the ambiguity in the corpus: *bat* not only shares a frame (b_t) with *boat* and *bite*, but also, in the C15 dataset, with *put*, *pad* and *bad* (b/p_d/t), and in the C6 dataset, with *dog* and *kite*, among many others (STOP_STOP). Table 1 shows the percentage of types and tokens that are ambiguous in each dataset, that is, words in frames that match multiple wordtypes. Note that we always evaluate against the *gold* word identities, even when these are not distinguished in the model's input. These datasets are intended to evaluate the degree of reliance on consonant information in the LD and TLD models, and to what extent the topics in the TLD model can replace this information.

6.4 Topics

The input to the TLD model includes a distribution over topics for each situation, which we infer in advance from the full Brent corpus (not only the C1 subset) using LDA. Each transcript in the Brent corpus captures about 75 minutes of parent-child interaction, and thus multiple situations will be included in each file. The transcripts do not delimit situations, so we do this somewhat arbitrarily by splitting each transcript after 50 CDS utterances, resulting in 203 situations for the Brent C1 dataset. As well as function words, we also remove the five most frequent content words (*be, go, get, want, come*). On average, situations are only 59 words long, reflecting the relative lack of content words in CDS utterances.

Dataset	C24	C15	C6
Input Types	1487	1426	1203
Frames	1259	1078	702
Ambig Types %	27.2	42.0	80.4
Ambig Tokens %	41.3	56.9	77.2

Table 1: Corpus statistics showing the increasing amount of ambiguity as consonant categories are merged. Input types are the number of word types with distinct input representations (as opposed to gold orthographic word types, of which there are 1497). Ambiguous types and tokens are those with frames that match multiple (orthographic) word types.

We infer 50 topics for this set of situations using the `mallet` toolkit (McCallum, 2002). Hyperparameters are inferred, which leads to a dominant topic that includes mainly light verbs (*have, let, see, do*). The other topics are less frequent but capture stronger semantic meaning (e.g. *yummy, peach, cookie, daddy, bib* in one topic, *shoe, let, put, hat, pants* in another). The word-topic assignments are used to calculate unsmoothed situation-topic distributions $\boldsymbol{\theta}$ used by the TLD model.

6.5 Evaluation

We evaluate against adult categories, i.e., the 'gold-standard', since all learners of a language eventually converge on similar categories. (Since our model is not a model of the learning process, we do not compare the infant learning process to the learning algorithm.) We evaluate both the inferred phonetic categories and words using the clustering evaluation measure V-Measure (VM; Rosenberg and Hirschberg, 2007).[6] VM is the harmonic mean of two components, similar to F-score, where the components (VC and VH) are measures of cross entropy between the gold and model categorization.

[6]Other clustering measures, such as 1-1 matching and pairwise precision and recall (accuracy and completeness) showed the same trends, but VM has been demonstrated to be the most stable measure when comparing solutions with varying numbers of clusters (Christodoulopoulos et al., 2010).

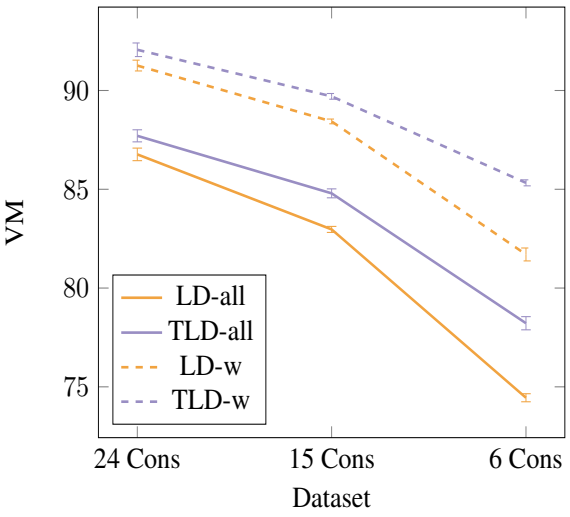

Figure 3: Vowel evaluation. 'all' refers to datasets with vowels synthesized from all speakers, 'w' to datasets with vowels synthesized from adult female speakers' vowels. The bars show a 95% Confidence Interval based on 5 runs. IGMM-all results in a VM score of 53.9 (CI=0.5); IGMM-w has a VM score of 65.0 (CI=0.2), not shown.

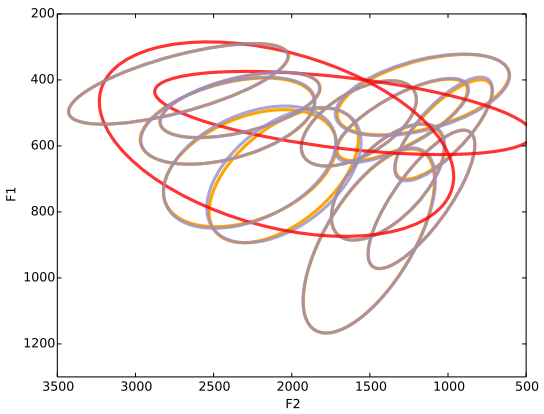

Figure 4: Vowels found by the TLD model; supervowels are indicated in red. The gold-standard vowels are shown in gold in the background but are mostly overlapped by the inferred categories.

For vowels, VM measures how well the inferred phonetic categorizations match the gold categories; for lexemes, it measures whether tokens have been assigned to the same lexemes both by the model and the gold standard. Words are evaluated against gold orthography, so homophones, e.g. *hole* and *whole*, are distinct gold words.

6.6 Results

We compare all three models—TLD, LD, and IGMM—on the vowel categorization task, and TLD and LD on the lexical categorization task (since IGMM does not infer a lexicon). The datasets correspond to two sets of conditions: firstly, either using vowel categories synthesized from all speakers or only adult female speakers, and secondly, varying the coarseness of the observed consonant categories. Each condition (model, vowel speakers, consonant set) is run five times, using 1500 iterations of Gibbs sampling with hyperparameter sampling. Overall, we find that TLD outperforms the other models in both tasks, across all conditions.

Vowel categorization results are shown in Figure 3. IGMM performs substantially worse than both TLD and LD, with scores more than 30 points lower than the best results for these models, clearly showing the value of the protolexicon and replicating the results found by Feldman et al. (2013a) on this dataset. Furthermore, TLD consistently outperforms the LD model, finding better phonetic categories, both for vowels generated from the combined categories of all speakers ('all') and vowels generated from adult female speakers only ('w'), although the latter are clearly much easier for both models to learn. Both models perform less well when the consonant frames provide less information, but the TLD model performance degrades less than the LD performance.

Both the TLD and the LD models find 'supervowel' categories, which cover multiple vowel categories and are used to merge minimal pairs into a single lexical item. Figure 4 shows example vowel categories inferred by the TLD model, including two supervowels. The TLD supervowels are used much less frequently than the supervowels found by the LD model, containing, on average, only two-thirds as many tokens.

Figure 5 shows that TLD also outperforms LD on the lexeme/word categorization task. Again performance decreases as the consonant categories become coarser, but the additional semantic information in the TLD model compensates for the lack of consonant information. In the individual components of VM, TLD and LD have similar VC ("recall"), but TLD has higher VH ("precision"), demonstrating that the semantic information given by the topics can separate potentially ambiguous words, as hypothesized.

Overall, the contextual semantic information

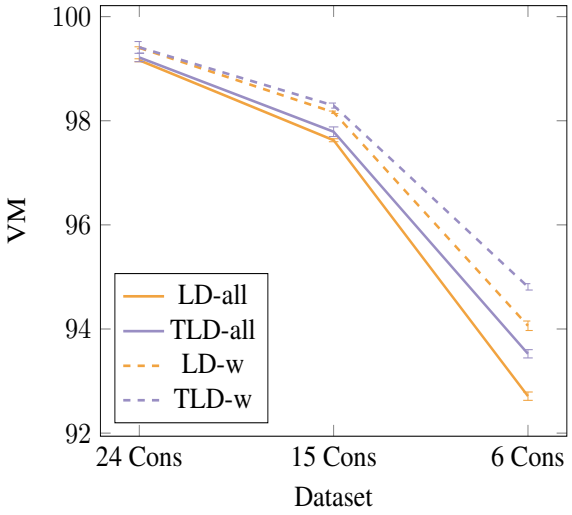

Figure 5: Lexeme evaluation. 'all' refers to datasets with vowels synthesized from all speakers, 'w' to datasets with vowels synthesized from adult female speakers' vowels.

added in the TLD model leads to both better phonetic categorization and to a better protolexicon, especially when the input is noisier, using degraded consonants. Since infants are not likely to have perfect knowledge of phonetic categories at this stage, semantic information is a potentially rich source of information that could be drawn upon to offset noise from other domains. The form of the semantic information added in the TLD model is itself quite weak, so the improvements shown here are in line with what infant learners could achieve.

7 Conclusion

Language acquisition is a complex task, in which many heterogeneous sources of information may be useful. In this paper, we investigated whether contextual semantic information could be of help when learning phonetic categories. We found that this contextual information can improve phonetic learning performance considerably, especially in situations where there is a high degree of phonetic ambiguity in the word-forms that learners hear. This suggests that previous models that have ignored semantic information may have underestimated the information that is available to infants. Our model illustrates one way in which language learners might harness the rich information that is present in the world without first needing to acquire a full inventory of word meanings.

The contextual semantic information that the TLD model tracks is similar to that potentially used in other linguistic learning tasks. Theories of cross-situational word learning (Smith and Yu, 2008; Yu and Smith, 2007) assume that sensitivity to situational co-occurrences between words and non-linguistic contexts is a precursor to learning the meanings of individual words. Under this view, contextual semantics is available to infants well before they have acquired large numbers of semantic minimal pairs. However, recent experimental evidence indicates that learners do not always retain detailed information about the referents that are present in a scene when they hear a word (Medina et al., 2011; Trueswell et al., 2013). This evidence poses a direct challenge to theories of cross-situational word learning. Our account does not necessarily require learners to track co-occurrences between words and individual objects, but instead focuses on more abstract information about salient events and topics in the environment; it will be important to investigate to what extent infants encode this information and use it in phonetic learning.

Regardless of the specific way in which infants encode semantic information, our method of adding this information by using LDA topics from transcript data was shown to be effective. This method is practical because it can approximate semantic information without relying on extensive manual annotation.

The LD model extended the phonetic categorization task by adding word contexts; the TLD model presented here goes even further, adding larger situational contexts. Both forms of top-down information help the low-level task of classifying acoustic signals into phonetic categories, furthering a holistic view of language learning with interaction across multiple levels.

Acknowledgments

This work was supported by EPSRC grant EP/H050442/1 and a James S. McDonnell Foundation Scholar Award to the final author.

References

Frans Adriaans and Daniel Swingley. Distributional learning of vowel categories is supported by prosody in infant-directed speech. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci)*, 2012.

E. Bergelson and D. Swingley. At 6-9 months, human infants know the meanings of many

common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258, Feb 2012.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In *Advances in Neural Information Processing Systems 16*, 2003.

Michael R. Brent and Jeffrey M. Siskind. The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44, 2001.

Christos Christodoulopoulos, Sharon Goldwater, and Mark Steedman. Two decades of unsupervised POS induction: How far have we come? In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 575–584, Cambridge, MA, October 2010. Association for Computational Linguistics.

Bart de Boer and Patricia K. Kuhl. Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, 4(4): 129, 2003.

Brian Dillon, Ewan Dunbar, and William Idsardi. A single-stage approach to learning phonological categories: Insights from Inuktitut. *Cognitive Science*, 37(2):344–377, Mar 2013.

Micha Elsner, Sharon Goldwater, Naomi Feldman, and Frank Wood. A cognitive model of early lexical acquisition with phonetic variability. In *Proceedings of the 18th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

Naomi H. Feldman, Thomas L. Griffiths, Sharon Goldwater, and James L. Morgan. A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 2013a.

Naomi H. Feldman, Emily B. Myers, Katherine S. White, Thomas L. Griffiths, and James L. Morgan. Word-level information influences phonetic learning in adults and infants. *Cognition*, 127(3): 427–438, 2013b.

Abdellah Fourtassi and Emmanuel Dupoux. A rudimentary lexicon and semantics help bootstrap phoneme acquisition. Submitted.

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5): 578–585, 2009.

Manuela Friedrich and Angela D. Friederici. Word learning in 6-month-olds: Fast encoding—weak retention. *Journal of Cognitive Neuroscience*, 23 (11):3228–3240, Nov 2011.

Lakshmi J. Gogate and Lorraine E. Bahrick. Intersensory redundancy and 7-month-old infants' memory for arbitrary syllable-object relations. *Infancy*, 2(2):219–231, Apr 2001.

J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5 Pt 1):3099–3111, May 1995.

P. W. Jusczyk and Elizabeth A. Hohne. Infants' memory for spoken words. *Science*, 277(5334): 1984–1986, Sep 1997.

Patricia K. Kuhl, Karen A. Williams, Francisco Lacerda, Kenneth N. Stevens, and Bjorn Lindblom. Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608, 1992.

Brian MacWhinney. *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, 2000.

D. R. Mandel, P. W. Jusczyk, and D. B. Pisoni. Infants' recognition of the sound patterns of their own names. *Psychological Science*, 6(5):314–317, Sep 1995.

Nivedita Mani and Kim Plunkett. Twelve-month-olds know their cups from their keps and tups. *Infancy*, 15(5):445470, Sep 2010.

Jessica Maye, Daniel J. Weiss, and Richard N. Aslin. Statistical phonetic learning in infants: facilitation and feature generalization. *Developmental Science*, 11(1):122–134, Jan 2008.

Jessica Maye, Janet F Werker, and LouAnn Gerken. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111, Jan 2002.

Andrew McCallum. MALLET: A machine learning for language toolkit, 2002.

Bob McMurray, Richard N. Aslin, and Joseph C. Toscano. Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12(3):369–378, May 2009.

Tamara Nicol Medina, Jesse Snedeker, John C. Trueswell, and Lila R. Gleitman. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019, 2011.

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

Linda Polka and Janet F. Werker. Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, 20(2):421–435, 1994.

Carl Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 13*, 2000.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2007.

Brandon C. Roy, Michael C. Frank, and Deb Roy. Relating activity contexts to early word learning in dense longitudinal data. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci)*, 2012.

Rushen Shi and Janet F. Werker. The basis of preference for lexical words in 6-month-old infants. *Developmental Science*, 6(5):484–488, 2003.

M. Shukla, K. S. White, and R. N. Aslin. Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, 108 (15):6038–6043, Apr 2011.

Linda B. Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.

Christine L. Stager and Janet F. Werker. Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388: 381–382, 1997.

D. Swingley. Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3617–3632, Nov 2009.

Yee Whye Teh. A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 985 – 992, Sydney, 2006.

Tuomas Teinonen, Richard N. Aslin, Paavo Alku, and Gergely Csibra. Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, 108:850–855, 2008.

Erik D. Thiessen. The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, 56(1):16–34, Jan 2007.

R. Tincoff and P. W. Jusczyk. Some beginnings of word comprehension in 6-month-olds. *Psychological Science*, 10(2):172–175, Mar 1999.

Ruth Tincoff and Peter W. Jusczyk. Six-month-olds comprehend words that refer to parts of the body. *Infancy*, 17(4):432444, Jul 2012.

N. S. Trubetzkoy. *Grundzüge der Phonologie*. Vandenhoeck und Ruprecht, Göttingen, 1939.

John C. Trueswell, Tamara Nicol Medina, Alon Hafri, and Lila R. Gleitman. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66:126–156, 2013.

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

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

H. Henny Yeung and Janet F. Werker. Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, 113(2): 234–243, Nov 2009.

Chen Yu and Linda B. Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420, 2007.