A Joint Learning Model of Word Segmentation, Lexical Acquisition, and Phonetic Variability

A Joint Learning Model of Word Segmentation, Lexical Acquisition, and Phonetic Variability

Micha Elsner
melsner0@gmail.com
Dept. of Linguistics
The Ohio State University

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

Naomi H. Feldman
nhf@umd.edu
Dept. of Linguistics
University of Maryland

Frank Wood
fwood@robots.ox.ac.uk
Dept. of Engineering
University of Oxford

Abstract

We present a cognitive model of early lexical acquisition which jointly performs word segmentation and learns an explicit model of phonetic variation. We define the model as a Bayesian noisy channel; we sample segmentations and word forms simultaneously from the posterior, using beam sampling to control the size of the search space. Compared to a pipelined approach in which segmentation is performed first, our model is qualitatively more similar to human learners. On data with variable pronunciations, the pipelined approach learns to treat syllables or morphemes as words. In contrast, our joint model, like infant learners, tends to learn multiword collocations. We also conduct analyses of the phonetic variations that the model learns to accept and its patterns of word recognition errors, and relate these to developmental evidence.

1 Introduction

By the end of their first year, infants have acquired many of the basic elements of their native language. Their sensitivity to phonetic contrasts has become language-specific (Werker and Tees, 1984), and they have begun detecting words in fluent speech (Jusczyk and Aslin, 1995; Jusczyk et al., 1999) and learning word meanings (Bergelson and Swingley, 2012). These developmental cooccurrences lead some researchers to propose that phonetic and word learning occur jointly, each one informing the other (Swingley, 2009; Feldman et al., 2013). Previous computational models capture some aspects of this joint learning problem, but typically simplify the problem considerably, either by assuming an unrealistic degree of phonetic regularity for word segmentation (Goldwater et al., 2009) or assuming pre-segmented input for phonetic and lexical acquisition (Feldman et al., 2009; Feldman et al., in press; Elsner et al., 2012). This paper presents, to our knowledge, the first broad-coverage model that learns to segment phonetically variable input into words, while simultaneously learning an explicit model of phonetic variation that allows it to cluster together segmented tokens with different phonetic realizations (e.g., *[ju]* and *[jɪ]*) into lexical items (*/ju/*).

We base our model on the Bayesian word segmentation model of Goldwater et al. (2009) (henceforth GGJ), using a noisy-channel setup where phonetic variation is introduced by a finite-state transducer (Neubig et al., 2010; Elsner et al., 2012). This integrated model allows us to examine how solving the word segmentation problem should affect infants' strategies for learning about phonetic variability and how phonetic learning can allow word segmentation to proceed in ways that mimic the idealized input used in previous models.

In particular, although the GGJ model achieves high segmentation accuracy on phonemic (non-variable) input and makes errors that are qualitatively similar to human learners (tending to undersegment the input), its accuracy drops considerably on phonetically noisy data and it tends to oversegment rather than undersegment. Here, we demonstrate that when the model is augmented to account for phonetic variability, it is able to learn common phonetic changes

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 42–54,
Seattle, Washington, USA, 18-21 October 2013. ©2013 Association for Computational Linguistics

and by doing so, its accuracy improves and its errors return to the more human-like undersegmentation pattern. In addition, we find small improvements in lexicon accuracy over a pipeline model that segments first and then performs lexical-phonetic learning (Elsner et al., 2012). We analyze the model's phonetic and lexical representations in detail, drawing comparisons to experimental results on adult and infant speech processing. Taken together, our results support the idea that a Bayesian model that jointly performs word segmentation and phonetic learning provides a plausible explanation for many aspects of early phonetic and word learning in infants.

2 Related Work

Nearly all computational models used to explore the problems addressed here have treated the learning tasks in isolation. Examples include models of word segmentation from phonemic input (Christiansen et al., 1998; Brent, 1999; Venkataraman, 2001; Swingley, 2005) or phonetic input (Fleck, 2008; Rytting, 2007; Daland and Pierrehumbert, 2011; Boruta et al., 2011), models of phonetic clustering (Vallabha et al., 2007; Varadarajan et al., 2008; Dupoux et al., 2011) and phonological rule learning (Peperkamp et al., 2006; Martin et al., 2013).

Elsner et al. (2012) present a model that is similar to ours, using a noisy channel model implemented with a finite-state transducer to learn about phonetic variability while clustering distinct tokens into lexical items. However (like the earlier lexical-phonetic learning model of Feldman et al. (2009; in press)) their model assumes known word boundaries, so to perform both segmentation and lexical-phonetic learning, they use a pipeline that first segments using GGJ and then applies their model to the results.

Neubig et al. (2010) also present a transducer-based noisy channel model that performs joint inference on two out of the three tasks we consider here; their model assumes fixed probabilities for phonetic changes (the noise model) and jointly infers the word segmentation and lexical items, as in our 'oracle' model below (though unlike our system their model learns from phone lattices rather than a single transcription). They evaluate only on phone recognition, not scoring the inferred lexical items.

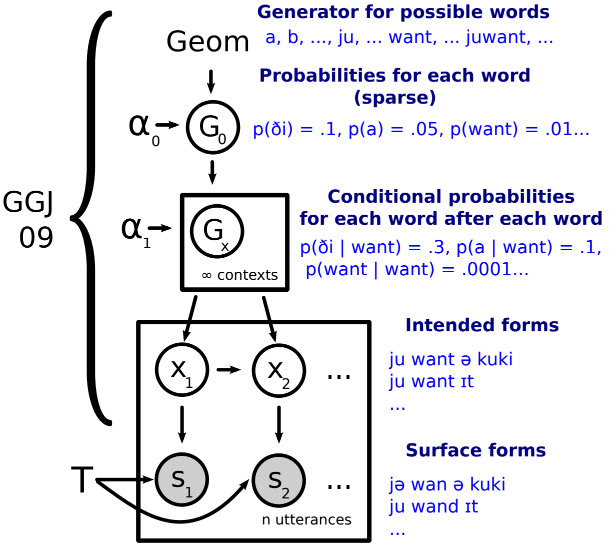
Generator for possible words
Geom
a, b, ..., ju, ... want, ... juwant, ...
Probabilities for each word (sparse)
α0
G0
p(ði) = .1, p(a) = .05, p(want) = .01...
GGJ 09
Conditional probabilities for each word after each word
α1
Gx
∞ contexts
p(ði | want) = .3, p(a | want) = .1,
p(want | want) = .0001...
Intended forms
X1
X2
...
ju want ə kuki
ju want ɪt
...
Surface forms
T
S1
S2
...
n utterances
jə wan ə kuki
ju wand ɪt
...

Figure 1: The graphical model for our system (Eq. 1-4). Note that the s_i are not distinct observations; they are concatenated together into a continuous sequence of characters which constitute the observations.

Recently, Börschinger et al. (2013) did present a joint learner for segmentation, phonetic learning, and lexical clustering, but the model and inference are tailored to investigate word-final /t/-deletion, rather than aiming for a broad coverage system as we do.

3 Model

We follow several previous models of lexical acquisition in adopting a Bayesian noisy channel framework (Eq. 1-4; Fig. 1). The model has two components: a *source* distribution $P(X)$ over utterances without phonetic variability X, i.e., *intended forms* (Elsner et al., 2012) and a *channel* or *noise* distribution $T(S|X)$ that translates them into the observed surface forms S. The boundaries between surface forms are then deterministically removed so that the actual observations are just the unsegmented string of characters in the surface forms.

$$G_0|\alpha_0, p_{stop} \sim DP(\alpha_0, Geom(p_{stop})) \quad (1)$$

$$G_x|G_0, \alpha_1 \sim DP(\alpha_1, G_0) \quad (2)$$

$$X_i|X_{i-1} \sim G_{X_{i-1}} \quad (3)$$

$$S|X;\theta \sim T(S|X;\theta) \quad (4)$$

The source model is an exact copy of GGJ[1]: to generate the intended-form word sequences X, we

[1] We use their best reported parameter values: $\alpha_0 = 3000, \alpha_1 = 100, p_{stop} = .2$ and for unigrams, $\alpha_0 = 20$.

sample a random language model from a hierarchical Dirichlet process (Teh et al., 2006) with character strings as atoms. To do so, we first draw a unigram distribution G_0 from a Dirichlet process prior whose base distribution generates intended form word strings by drawing each phone in turn until the stop character is drawn (with probability p_{stop}). Then, for each possible context word x, we draw a conditional distribution on words following that context $G_x = P(X_i = \bullet | X_{i-1} = x)$ using G_0 as a prior. Finally, we sample word sequences $x_1 \ldots x_n$ from the bigram model.

The channel model is a finite transducer with parameters θ which independently rewrites single characters from the intended string into characters of the surface string. We use MAP point estimates of these parameters; single characters (without n-gram context) are used for computational efficiency. Also for efficiency, the transducer can insert characters into the surface string, but cannot delete characters from the intended string. As in several previous phonological models (Dreyer et al., 2008; Hayes and Wilson, 2008), the probabilities are learned using a feature-based log-linear model. For features, we use all the unigram features from Elsner et al. (2012), which check faithfulness to voicing, place and manner of articulation (for example, for $k \rightarrow g$, active features are *faith-manner*, *faith-place*, *output-g* and *voiceless-to-voiced*).

Below, we present two methods for learning the transducer parameters θ. The *oracle* transducer is estimated using the gold-standard word segmentations and intended forms for the dataset; it represents the best possible approximation under our model of the actual phonetics of the dataset. We can also estimate the transducer using the *EM* algorithm. We first initialize a simple transducer by putting small weights on the faithfulness features to encourage phonologically plausible changes. With this initial model, we begin running the sampler used to learn word segmentations. After several hundred sampler iterations, we start re-estimating the transducer by maximum likelihood after each iteration. We regularize our estimates by adding 200 pseudocounts for the rewrite $x \rightarrow x$ during training (rather than regularizing the weights for particular features). We also show *segment only* results for a model without the transducer component (i.e., $S = X$); this recovers the GGJ baseline.

4 Inference

Inference for this model is complicated for two reasons. First, the hypothesis space is extremely large. Since we allow the input string to be probabilistically lengthened, we cannot be sure how long it is, nor which characters it contains. Second, our hypotheses about nearby characters are highly correlated due to lexical effects. When deciding how to interpret *[wənt]*, if we posit that the intended vowel is /ʌ/, the word is likely to be /wʌn/ "one" and the next word begins with /t/; if instead we posit that the vowel is /ɔ/, the word is probably /wɔnt/ "want". Thus, inference methods that change only one character at a time are unlikely to mix well. Since they cannot simultaneously change the vowel and resegment the /t/, they must pass through a low-probability intermediate state to get from one state to the other, so will tend to get stuck in a bad local minimum. A Gibbs sampler which inserts or deletes a single segment boundary in each step (Goldwater et al., 2009) suffers from this problem.

Mochihashi et al. (2009) describe an inference method with higher mobility: a block sampler for the GGJ model that samples from the posterior over analyses of a whole utterance at once. This method encodes the model as a large HMM, using dynamic programming to select an analysis. We encode our own model in the same way, constructing the HMM and composing it with the transducer (Mohri, 2004) to form a larger finite-state machine which is still amenable to forward-backward sampling.

4.1 Finite-state encoding

Following Mochihashi et al. (2009) and Neubig et al. (2010), we can write the original GGJ model as a Hidden Semi-Markov model. States in the HMM, written ST:$[w][C]$, are labeled with the previous word w and the sequence of characters C which have so far been incorporated into the current word. To produce a word boundary, we transition from ST:$[w][C]$ to ST:$[C][]$ with probability $P(x_i = C | x_{i-1} = w)$. We can also add the next character s to the current word, transitioning from ST:$[w][C]$ to ST:$[w][C:s]$, at no cost (since the full cost of the word is paid at its boundary, there

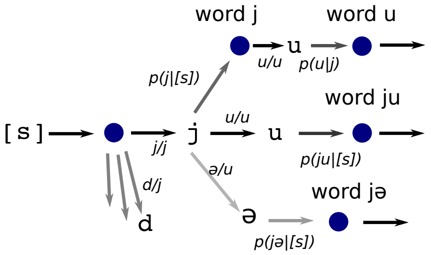

Figure 2: A fragment of the composed finite-state machine for word segmentation and character replacement for the surface string *ju*. The start state [s] is followed by a word boundary (filled circle); the next intended character is probably *j* but can be *d* or others with lower probability. After *j* can be a word boundary (forming the intended word *j*), or another character such as *u*, ə or other (not shown) alternatives.

is no cost for the individual characters)[2].

In addition to analyses using known words, we can also encode the uniform-geometric prior over unknown words using a finite-state machine. We can choose to select a word from the prior by transitioning to a state `ST:[`$Geom$`][]` with probability $P(\text{new word}|x_{i-1} = w)$ immediately after a word boundary. While in $Geom$, we can transition to a new $Geom$ state and produce any character with uniform probability $P(c) = (1 - P_{stop})\frac{1}{|C|}$; otherwise, we can end the word, transitioning to `ST:[`$unk.word$`][]`, with probability P_{stop}.

This construction is also approximate; it ignores the possibility that the prior will generate a known word w, in which case our final transition ought to be to `ST:[`w`][]` instead of `ST:[`$unk.word$`][]`. This approximation means we do not need to add context to the $Geom$ state to remember the sequence of characters it produced, which allows us to keep only a single $Geom$ state on the chart at each timestep.

When we compose this model with the channel model, the number of states expands. Each state must now keep track of the previous word, what intended characters C have been posited and what surface characters S have been recognized, `ST:[`w`][`C`][`S`]`. To recognize the current word, we transition to `ST:[`C`][][]` with probability $P(x_i = C|x_{i-1} = w)$. To parse a new surface character s by positing intended character x (note that x might be ϵ), we transition to `ST:[`w`][`$C : x$`][`$S : s$`]` with probability $T(s|x)$. (As above, we pay no cost for our choice of x, which is paid for when we recognize the word; however, we must pay for s.) For efficiency, we do not allow the G_0 states to hypothesize different surface and intended characters, so when we initially propose an unknown word, it must surface as itself.[3]

4.2 Beam sampler

This machine has too many states to fully fill the chart before backward sampling, so we restrict the set of trajectories under consideration using beam sampling (Van Gael et al., 2008) and simulated annealing.

The beam sampler is closely related to the standard beam search technique, which uses a probability cutoff to discard parts of the FST which are unlikely to figure in the eventual solution. Unlike conventional beam search, the sampler explores using stochastic cutoffs, so that all trajectories are explored, but most of the bad ones are explored infrequently, leading to higher efficiency.

We design our beam sampler to restrict the set of potential intended characters at each timestep. In particular, given a stream of input characters $S = s_1 \dots s_n$, we introduce a set of auxiliary cutoff variables $U = u_1 \dots u_n$. The u_i variables represent limits on the probability of the emission of surface character s_i; we exclude any hypothesized x_i whose probability of generating s_i, $T(s_i|x_i)$, is less than u_i. To create a beam sampling scheme, we must devise a distribution for U given a state sequence Q (as discussed above, the sequence of states encodes the intended character sequence and the segmentation of the surface string), $P_u(U|Q)$ and then incorporate the probability of U into the forward messages.

If q_i is the state in Q at which s_i is generated, and x_i the corresponding intended character, we require that $P_u < T(s_i|x_i)$; that is, the cutoffs must not exclude any states in the sequence Q. We define P_u

[2] Though not mentioned by Mochihashi et al. (2009) or Neubig et al. (2010), this construction is not exact, since transitions in a Bayesian HMM are exchangeable but not independent (Beal et al., 2001): if a word occurs twice in an utterance, its probability is slightly higher the second time. For single utterances, this bias is small and easy to correct for using a Metropolis-Hastings acceptance check (Börschinger and Johnson, 2012) using the path probability from the HMM as the proposal.

[3] Again, this approximation is corrected for by the Metropolis-Hastings step.

as a λ-mixture of two distributions:

$$P_u(u|s_i, x_i) = \lambda U[0, min(.05, T(s_i|x_i))] + (1-\lambda)T(s_i|x_i)Beta(5, 1e-5)$$

The former distribution is quite unrestrictive, while the latter prefers to prune away nearly all the states. Thus, for most characters in the string, we do not permit radical changes, while for a fraction, we do.

We follow Huggins and Wood (2013), who extended Van Gael et al. (2008) to the case of a non-uniform P_u, to define our forward message α as:

$$\begin{aligned} \alpha(q_i, i) &\propto P(q_i, S_{0..i}, U_{0..i}) \qquad (5) \\ &= \sum_{q_{i-1}} P_u(u_i|s_i, x_i)T(s_i|x_i)\alpha(q_{i-1}, i-1) \end{aligned}$$

This is the standard HMM forward message, augmented with the probability of u. Since $P_u(\cdot|s_i, x_i)$ is required to be less than $T(s_i|x_i)$, it will be 0 whenever $T(s_i|x_i) < u$; this is how the u variables function as cutoffs. In practice, we use the u variables to filter the lexical items that begin at each position i in advance, using a simple 0/1 edit distance Markov model which runs faster than our full model. (For example, we can quickly check if the current U allows *want* as the intended form for *wɔlk* at i; if not, we can avoid constructing the prefix `ST:` [x_{i-1}] [wa] [wɔ] since the continuation will fail.)

The algorithm's speed depends on the size and uncertainty of the inferred LM: large numbers of plausible words mean more states to explore. When inference starts, and the system is highly uncertain about word boundaries, it is therefore reasonable to limit the exploration of the character sequence. We do so by annealing in two ways: as in Goldwater et al. (2009), we raise $P(X)$ (Eq. 3) to a power t which increases linearly from .3. To sample from the posterior, we would want to end with $t = 1$, but as in previous noisy-channel models (Elsner et al., 2012; Bahl et al., 1980) we get better results when we emphasize the LM at the expense of the channel and so end at $t = 2$. Meanwhile, as t rises and we explore fewer implausible lexical sequences, we can explore the character sequence more. We begin by setting the λ interpolation parameter of P_u to 0 to minimize exploration and increase it linearly to .3 (allowing the system to change about a third of the characters on each sweep). This is similar to the scheme for altering P_u in Huggins and Wood (2013).

4.3 Dataset and metrics

We use the corpus released by Elsner et al. (2012), which contains 9790 child-directed English utterances originally from the Bernstein-Ratner corpus (Bernstein-Ratner, 1987) and later transcribed phonemically (Brent, 1999). This standard word segmentation dataset was modified by Elsner et al. (2012) to include phonetic variation by assigning each token a pronunciation independently selected from the empirical distribution of pronunciations of that word type in the closely-transcribed Buckeye Speech Corpus (Pitt et al., 2007). Following previous work, we hold out the last 1790 utterances as unseen test data during development. In the results presented here, we run the model on all 9790 utterances but score only these 1790. We average results over 5 runs of the model with different random seeds.

We use standard metrics for segmentation and lexicon recovery. For segmentation, we report precision, recall and F-score for word boundaries (*bds*), and for the positions of word tokens in the surface string (*srf*; both boundaries must be correct).

For normalization of the pronunciation variation, we follow Elsner et al. (2012) in measuring how well the system clusters together variant pronunciations of the same lexical item, without insisting that the intended form the system proposes for them match the one in our corpus. For example, if the system correctly clusters *[ju]* and *[jɪ]* together but assigns them the incorrect intended form */jɪ/*, we can still give credit to this cluster if it is the one that overlaps best with the gold-standard */ju/* cluster. To compute these scores, we find the optimal one-to-one mapping between our clusters of pronunciations and the true lexical entries, then report scores for mapped tokens (*mtk*; boundaries and mapping to gold standard cluster must be correct) and mapped types[4] (*mlx*).

[4] Elsner et al. (2012) calls the *mlx* metric *lexicon F*, which is possibly confusing. We map the clusters to a gold-standard lexicon (plus potentially some words that don't correspond to anything in the gold standard) and compute a type-level F-score on this lexicon.

	Prec	Rec	F-score
Pipeline (segment, then cluster): (Elsner et al., 2012)			
Bds	70.4	**93.5**	80.3
Srf	56.5	69.7	62.4
Mtk	44.2	**54.5**	48.8
Mlx	48.6	43.1	45.7
Bigram model, segment only			
Bds	73.9 (-0.6:0.7)	91.0 (-0.6:0.4)	**81.6** (-0.5:0.6)
Srf	60.8 (-0.7:1.1)	**70.8** (-0.8:0.9)	65.4 (-0.6:1.0)
Mtk	41.6 (-0.6:1.2)	48.4 (-0.5:1.2)	44.8 (-0.6:1.2)
Mlx	36.6 (-0.7:0.8)	**49.8** (-1.0:0.8)	42.2 (-0.9:0.8)
Unigram model, oracle transducer			
Bds	**81.4** (-0.8:0.4)	72.1 (-0.9:0.8)	76.4 (-0.5:0.7)
Srf	63.6 (-1.0:1.1)	58.5 (-1.2:1.2)	60.9 (-0.9:1.2)
Mtk	46.8 (-1.0:1.1)	43.0 (-1.1:1.2)	44.8 (-1.0:1.2)
Mlx	**56.7** (-1.1:1.0)	47.6 (-1.4:0.8)	**51.7** (-1.2:0.8)
Bigram model, oracle transducer			
Bds	76.1 (-0.6:0.6)	83.8 (-0.9:1.0)	79.8 (-0.8:0.4)
Srf	62.2 (-0.9:1.0)	66.7 (-1.2:1.1)	64.4 (-1.1:0.8)
Mtk	47.2 (-0.7:0.9)	50.6 (-1.0:0.8)	48.8 (-0.8:0.7)
Mlx	40.1 (-1.0:1.2)	43.7 (-0.6:0.7)	41.8 (-0.8:0.6)
Bigram model, EM transducer			
Bds	80.1 (-0.5:0.8)	83.0 (-1.4:1.3)	81.5 (-0.5:0.7)
Srf	**66.1** (-0.8:1.4)	67.8 (-1.4:1.7)	**66.9** (-0.9:1.4)
Mtk	**49.0** (-0.9:0.7)	50.3 (-1.1:1.4)	**49.6** (-1.0:1.0)
Mlx	43.0 (-1.0:1.4)	49.5 (-1.5:1.1)	46.0 (-1.0:1.3)

Table 1: Mean segmentation (*bds*, *srf*) and normalization (*mtk*, *mlx*) scores on the test set over 5 runs. Parentheses show min and max scores as differences from the mean.

5 Results and discussion

In the following sections, we analyze how our model with variability compares to GGJ on noisy data. We give quantitative scores and also show that qualitative patterns of errors are often similar to those of human learners and listeners.

5.1 Clean versus variable input

We begin by evaluating our model as a word segmentation system. (Table 1 gives segmentation and normalization scores for various models and baselines on the 1790 test utterances.) We first confirm that our inference method is reasonable. The bigram model without variability ("segment only") should have the same segmentation performance as the standard `dpseg` implementation of GGJ. This is the case: `dpseg` has boundary F of 80.3 and token F of 62.4; we get 81.6 and 65.4. Thus, our sampler is finding good solutions, at least for the no-variability model.

We compare segmentation scores between the "segment only" system and the two bigram models with transducers ("oracle" and "EM"). While these systems all achieve similar segmentation scores, they do so in different ways. "Segment only" finds a solution with boundary precision 73.9% and boundary recall 91.0% for a total F of 81.6%. The low precision and high recall here indicate a tendency to oversegment; when the analysis of a given subsequence is unclear, the system prefers to chop it into small chunks. The bigram models which incorporate transducers score P: 76.1, R: 83.8 (oracle) and P: 80.1, R: 83.0 (EM), indicating that they prefer to find longer sequences (undersegment) more.

In previous experiments on datasets without variation, GGJ also has a strong tendency to undersegment the data (boundary P: 90.1, R: 80.3), which Goldwater et al. argue is rational behavior for an ideal learner seeking a parsimonious explanation for the data. Undersegmentation occurs especially when ignoring lexical context (a unigram model), but to some extent even in bigram models. Human learners also tend to learn collocations as single words (Peters, 1983; Tomasello, 2000), and the GGJ model has been shown to capture several other effects seen in laboratory segmentation tasks (Frank et al., 2010). Together, these findings support the idea that human learners may behave in important respects like the Bayesian ideal learners that Goldwater et al. presented.

However, experiments on data with variation have called these conclusions into question. In particular, GGJ has previously been shown to oversegment rather than undersegment as the input grows noisier (Fleck, 2008), and our results replicate this finding (oversegmentation for the "segment only" model). In addition, the GGJ bigram model, which achieves much higher segmentation accuracy than the unigram model on clean data, actually performs worse on very noisy data (Jansen et al., 2013). Infants are known to track statistical dependencies across words (Gómez and Maye, 2005), so it is worrisome that these dependencies hurt GGJ's segmentation accuracy when learning from noisy data.

Our results show that modeling phonetic variability reverses the problematic trends described above. Although the models with phonetic variability show similar overall segmentation accuracy on noisy data to the original GGJ model, the pattern of errors changes, with less oversegmentation and more undersegmentation. Thus, their qualitative performance on variable data resembles GGJ's on clean data, and therefore the behavior of human learners.

5.2 Phonetic variability

We next analyze the model's ability to normalize variations in the pronunciation of tokens, by inspecting the *mtk* score. The "segment only" baseline is predictably poor, F: 44.8. The pipeline model scores 48.8, and our oracle transducer model matches this exactly. The EM transducer scores better, F: 49.6. Although the confidence intervals overlap slightly, the EM system also outperforms the pipeline on the other F-measures; altogether, these results suggest at least a weak learning synergy (Johnson, 2008) between segmentation and phonetic learning.

It is interesting that EM can perform better than the oracle. However, EM is more conservative about which sound changes it will allow, and thus tends to avoid mistakes caused by the simplicity of the transducer model. Since the transducer works segment-by-segment, it can apply rare contextual variations out of context. EM benefits from not learning these variations to begin with.

We can also compare the bigram and unigram versions of the model. The unigram model is a reasonable segmenter, though not quite as good as the bigram model, with boundary F of 76.4 and token F of 60.9 (compared to 79.8 and 64.4 using the bigram model). However, it is not good at normalizing variation; its *mtk* score is comparable to the baseline at 44.8%[5]. Although bigram context is only moderately effective for telling where words are, the model seems heavily reliant on lexical context to decide *what* words it is hearing.

5.3 Error analysis

To gain more insight into the differing behavior of our model versus a pipelined system, we inspect the intended word strings X proposed by each one in detail. Below, we categorize the kinds of intended word strings that the model might propose to span a given gold-standard word token:

Correct Correctly segmented, mapped to the correct lexical item (e.g., gold intended */ju/*, surface segmentation *[ju]*, intended */ju/*)

	EM-learned	Segment only
Correct	49.88	47.61
Wrong form	17.96	23.73
Collocation	14.25	7.59
Split	8.26	15.18
One bound	7.11	15.18
Corr. colloc.	1.35	< 0.01
Other	0.75	0.22
Corr. split	0.43	0.66

Table 2: Distribution (%) of error types (see text) in a single run on the full dataset.

Wrong form Correctly segmented, mapped to the wrong lexical item (*/ju/*, surf. *[ju]*, int. */jɛs/*)

Colloc Missegmented as part of a sequence whose boundaries correspond to real word boundaries (*/ju•want/*, surf. *[juwant]*, int. */juwant/*)

Corr. colloc As above, but proposed lexical item maps to this word (*/ar•ju/*, surf. *[arjə]* int. */ju/*)

Split Missegmented with a word-internal boundary (*/dɔgiz/*, surf. *[dɔ•giz]*, int. */dɔ•giz/*)

Corr. split As above, but one proposed word maps correctly (*/dɔgi/*, surf. *[dɔg•i]*, int. */dɔgi•ə/*)

One boundary One boundary correct, the other wrong (*/ju•wa.../*, surf. *[juw]*, int. */juw/*)

Other Not a collocation, both boundaries are wrong (*/du•ju•wa.../*, surf. *[ujuw]*, int. */ujuw/*)

Table 2 shows the distribution over intended word strings proposed by the "segment only" baseline and the EM-learned transducer. Both systems propose a large number of correct forms, and the most common error category is "wrong form" (lexical error without segmentation error), an error which could potentially be repaired in a pipeline system. However, the remaining errors represent segmentation mistakes which a pipeline could not repair. Here the two systems behave quite differently. The EM-learned transducer analyses 14% of real tokens as parts of multiword collocations like "doyou"; in another 1.35%, the underlying content word is even correctly detected. The non-variable system, on the other hand, analyses 15% of real tokens by splitting them into pieces. Since infant learners tend to learn collocations, this supports our analysis that the model with variation better models human behavior.

[5] Elsner et al. (2012) show a similar result for a unigram version of their pipelined system.

EM	*ju*: 805, *duju*: 239, *juwan*: 88, *jɪ*: 58, *e˞ju*: 54, *judu*: 47, *jæ*: 39, *julʌk*: 39, *ʃu*: 30, *u*: 23, *ʒu*: 18, *j*: 17, *je˞*: 16, *tʃu*: 15, *aj*:15, *ðerjugo*: 12, *dʒu*: 12
GGJ	*ju*: 498, *jɪ*: 280, *jə*: 165, *ji*: 119, *duju*: 106, *dujɪ*: 44, *kɪnju*: 39, *i*: 32, *u*: 29, *kɪnjɪ*: 29, *julʌk*: 24, *juwan*: 23, *j*: 22, *ʃu*: 19, *jʊ*: 18, *e˞ju*: 18, *ɪ*:16, *ʒu*: 15, *dʒ•u*: 13, *jɛ*: 12, *ʃɪ*: 11, *θæŋkju*: 11

Table 3: Forms proposed with frequency > 10 for gold-standard tokens of "you" in one sample from EM-transducer and segment-only (GGJ) system.

To illustrate this behavior anecdotally, we present the distribution of intended word strings spanning tokens whose gold intended form is */ju/* "you" (Table 3). The EM-learned solution proposes 805 tokens of */ju/*, which is the correct analysis[6]; the "segment only" system instead finds varying forms like */jɪ/*, */jæ/* etc. This is unsurprising and could be repaired by a suitable pipelined system. However, the EM system also proposes 239 instances of "doyou", 88 instances of "youwant", 54 instances of "areyou" and several other collocations. The "segment only" system finds some of these collocations, split into different versions: for instance 106 instances of */duju/* and 44 of */dujɪ/*. In a pipelined system, we could combine these variants to find 150 instances— but this is still 89 instances short of the 239 found when allowing for variability. The same pattern holds for "youlike" and "youwant". Because the non-variable system must learn each variant separately, it learns only the most common instances of these long collocations, and analyzes infrequent variants differently.

We also perform this analysis specifically for words beginning with vowels. Infants show a delay in their ability to segment these words from continuous speech (Mattys and Jusczyk, 2001; Nazzi et al., 2005; Seidl and Johnson, 2008), and Seidl and Johnson (2008) suggest a perceptual explanation— initial vowels can be hard to hear and often exhibit variation due to coarticulation or resyllabification. Although our dataset does not contain coarticulation as such, it should show this pattern of greater variation, which we hypothesize might lead to difficulty in segmenting and recognizing vowel-initial words.

The model's behavior is consistent with this hypothesis (Table 4). Both the "segment only" and EM transducer models find approximately the same proportion of vowel-initial tokens, and both systems do somewhat better on consonant-initial words than vowel-initial words. The advantage is stronger for the transducer model, which gets only 41.5% of vowel-initial tokens correct as opposed to 52.1% of consonant-initial words. It proposes more collocations for vowel-initial words (19.2%) than for consonants (12.5%). In cases where they do not propose a collocation, both systems are somewhat more likely to find the right boundary of a vowel-initial token than the left boundary (although again this difference is larger for the EM system); this suggests that the problem is indeed caused by the initial segment.

Segment only	Vow. init	Cons. init
Correct	47.5	51.7
Wrong form	18.6	15.7
Collocation	14.6	12.2
Split	6.2	10.8
Right bd. corr.	5.8	3.6
Left bd. corr.	4.6	3.8
EM transducer	Vow. init	Cons. init
Correct	41.5	52.1
Wrong form	20.4	17.3
Collocation	19.2	12.5
Split	5.2	9.1
Right bd. corr.	6.2	2.7
Left bd. corr.	2.7	3.1

Table 4: Most common error types (%; see text) for intended forms beginning with vowels or consonants. Rare error types are not shown. "One bound" errors are split up by which boundary is correct.

5.4 Phonetic Learning

We next compare phonetic variations learned by the model to characteristics of infant speech perception. Infants show an asymmetry between consonants and vowels, losing sensitivity to non-native vowel contrasts by eight months (Kuhl et al., 1992; Bosch and Sebastián-Gallés, 2003) but to non-native consonant contrasts only by 10-12 months (Werker and Tees, 1984). The observed ordering is somewhat puzzling when one considers the availability for distributional information (Maye et al., 2002), which is much stronger for stop consonants than for vowels (Lisker and Abramson, 1964; Peterson and Barney, 1952). Infants are also conservative in generalizing across phonetic variability, showing a delayed abil-

[6] Not all the variants are merged, however. *jɪ*, *jæ*, *ʃu* etc. are still occasionally analyzed as separate lexical items.

ity to generalize across talkers, affects, and dialects. They have difficulty recognizing word tokens that are spoken by a different talker or in a different tone of voice until 11 months (Houston and Jusczyk, 2000; Singh et al., 2004), and the ability to adapt to unfamiliar dialects appears to develop even later, between 15 and 19 months (Best et al., 2009; Heugten and Johnson, in press; White and Aslin, 2011).

Similar to infants, our model shows both a vowel-consonant asymmetry and a reluctance to accept the full range of adult phonetic variability. Table 5 shows some segment-to-segment alternations learned in various transducers. The oracle learns a large amount of variation (*u* surfaces as itself only 68% of the time) involving many different segments, whereas EM is similar to infant learners in learning a more conservative solution with fewer alternations overall. Moreover, EM appears to identify patterns of variability in vowels before consonants. It learns a similar range of alternations for *u* as in the oracle, although it treats the sound as less variable than it actually is. It learns much less variability for consonants; it picks up the alternation of *ð* with *s* and *z*, but predicts that *ð* will surface as itself 91% of the time when the true figure is only 69%. And it fails to learn any meaningful alternations involving *k*. These results suggest that patterns of variability in vowels are more evident than patterns of variability in consonants when infants are beginning to solve the word segmentation problem.

To investigate the effect of data size on this conservativism, we ran the system on 1000 utterances instead of 9790. This leads to an even more conservative solution, with variations for *u* but none of the others (although *i* and *ð* still vary more than *k*).

System	x	top 4 outputs s							
Oracle	*u*	*u*	.68	ə	.05	*a*	.04	ʊ	.04
	i	*i*	.85	ɪ	.03	ə	.03	ɛ	.02
	ð	*ð*	.69	*s*	.07	[ϕ]	.07	*z*	.04
	k	*k*	.93	*d*	.02	*g*	.02		
	[ϕ]	*r*	.21	*h*	.11	*d*	.01	ə	.07
EM (full)	*u*	*u*	.75	ə	.08	ɪ	.04	ʊ	.03
	i	*i*	.90	ɪ	.04	ɛ	.02		
	ð	*ð*	.91	*s*	.03	*z*	0.1		
	k	*k*	.98						
	[ϕ]	ə	.32	ɪ	.14	*n*	.13	*t*	.13
EM (only 1000 utts)	*u*	*u*	.82	ɪ	.04	ə	.04	*a*	.02
	i	*i*	.97						
	ð	*ð*	.95						
	k	*k*	.99						
	[ϕ]	ə	.21	ɪ	.18	*t*	.12	*s*	.12

Table 5: Learned phonetic alternations: top 4 outputs s with $p > .001$ for inputs $x =$ uw (/*u*/), iy (/*i*/), dh (/*ð*/), k (/*k*/) and [ϕ], the null character. Outputs from [ϕ] are insertions. The oracle allows [ϕ] as an output (deletion) but for computational reasons, the model does not.

5.5 Segmentation and recognition errors

A particularly interesting set of errors are those that involve both a missegmentation and a simultaneous misrecognition, since the joint model is prone to such errors while the pipelined model is not. Relatively little is known about infants' misrecognitions of words in fluent speech, although it is clear that they find words in medial position harder (Plunkett, 2005; Seidl and Johnson, 2006). However, adults make missegmentation/misrecognition errors fairly often, especially when listening to noisy audio (Butterfield and Cutler, 1988). Such errors are more common when the misrecognized word belongs to a prosodically rare class and when the incorrectly hypothesized string contains frequent words (Cutler, 1990); phonetically ambiguous words are also more commonly recognized as the more frequent of two options (Connine et al., 1993). For the indefinite article "a" (often reduced to *[ə]*), lexical context is the main factor in deciding between ambiguous interpretations (Kim et al., 2012). In rapid speech, listeners have few phonetic cues to indicate whether it is present at all (Dilley and Pitt, 2010). Below, we analyze various misrecognitions made by our system (using the EM transducer), and find some similar effects.

The easiest cases to analyze are those with no missegmentation: the proposed boundaries are correct, and the proposed lexical entry corresponds to a real word[7], but not the correct one. Most of them correspond to homophones (Table 6).

Common cases with a missegmentation include *it* and *is*, *a* and *is*, *it's* and *is*, *who*, *who's* and *whose*, *that's* and *what's*, and *there* and *there's*. In general, these errors involve words which sometimes appear

[7] The one-to-one mapping can be misleading, as it may map a large cluster to a real word on the basis of one or two tokens if all other tokens correspond to a different word already used for another cluster. We manually filter out a few cases like this.

Actual	proposed	count
/*tu*/ “two”	/*tə*/ “to”	95
/*kin*/ “can”	/*kænt*/ “can’t”	67
/*ɛn*/ “and”	/*æn*/ “an”	61
/*hɪz*/ “his”	/*ɪz*/ “is”	57
/*ðə*/ “the”	/*ə*/ “ah”	51
/*wəts*/ “what’s”	/*wants*/ “wants”	40
/*wan*/ “want”	/*won*/ “won’t”	39
/*yu*/ “you”	/*yæ*/ “yeah”	39
/*fɚ*/ “for”	/*fɔr*/ “four”	30
/*hir*/ “here”	/*hil*/ “he’ll”	28

Table 6: Top ten errors involving confusion between real, correctly segmented words: the most common pronunciation of the actual token and its orthographic form, the same for the proposed token, and the frequency.

with a morpheme or clitic (which can easily be missegmented as part of something else), words which differ by one segment, and frequent function words which often appear in similar contexts. These tendencies match those shown by adult human listeners.

A particularly distinctive set of joint recognition and segmentation errors are those where an entire real token is treated as phonetic “noise”— that is, it is segmented along with an adjacent word, and the system clusters the whole sequence as a token of that word. The most common examples are “that’s a” identified as “that’s”, “have a” identified as “have”, “sees a” identified as “sees” and other examples involving “a”, a word which also frequently confuses humans (Kim et al., 2012; Dilley and Pitt, 2010). However, there are also instances of “who’s in” as “who’s”, “does it” as “does”, and “can you” as “can”.

6 Conclusion

We have presented a model that jointly infers word segmentation, lexical items, and a model of phonetic variability; we believe this is the first model to do so on a broad-coverage naturalistic corpus[8]. Our results show a small improvement in both segmentation and normalization over a pipeline model, providing evidence for a synergistic interaction between these learning tasks and supporting claims of interactive learning from the developmental literature on infants. We also reproduced several experimental findings; our results suggest that two vowel-consonant asymmetries, one from the word segmentation literature and another from the phonetic learning literature, are linked to the large variability in vowels found in natural corpora. The model’s correspondence with human behavioral results is by no means exact, but we believe these kinds of predictions might help guide future research on infant phonetic and word learning.

Acknowledgements

Thanks to Mary Beckman for comments. This work was supported by EPSRC grant EP/H050442/1 to the second author.

References

Lalit Bahl, Raimo Bakis, Frederick Jelinek, and Robert Mercer. 1980. Language-model/acoustic-channel-model balance mechanism. Technical disclosure bulletin Vol. 23, No. 7b, IBM, December.

Matthew J. Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. 2001. The infinite Hidden Markov Model. In *NIPS*, pages 577–584.

Elika Bergelson and Daniel Swingley. 2012. At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109:3253–3258.

Nan Bernstein-Ratner. 1987. The phonology of parent-child speech. In K. Nelson and A. van Kleeck, editors, *Children’s Language*, volume 6. Erlbaum, Hillsdale, NJ.

Catherine T. Best, Michael D. Tyler, Tiffany N. Gooding, Corey B. Orlando, and Chelsea A. Quann. 2009. Development of phonological constancy: Toddlers’ perception of native- and jamaican-accented words. *Psychological Science*, 20(5):539–542.

Benjamin Börschinger and Mark Johnson. 2012. Using rejuvenation to improve particle filtering for Bayesian word segmentation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–89, Jeju Island, Korea, July. Association for Computational Linguistics.

Benjamin Börschinger, Mark Johnson, and Katherine Demuth. 2013. A joint model of word segmentation and phonological variation for English word-final /t/-deletion. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August. Association for Computational Linguistics.

Luc Boruta, Sharon Peperkamp, Benoît Crabbé, and Emmanuel Dupoux. 2011. Testing the robustness of online word segmentation: Effects of linguistic diversity and

[8] Software is available from the ACL archive; updated versions may be posted at `https://bitbucket.org/melsner/beamseg`.

phonetic variation. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–9.

Laura Bosch and Núria Sebastián-Gallés. 2003. Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Language and Speech*, 46(2-3):217–243.

Michael R. Brent. 1999. An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, 34:71–105, February.

Sally Butterfield and Anne Cutler. 1988. Segmentation errors by human listeners: Evidence for a prosodic segmentation strategy. In *Proceedings of SPEECH '88: Seventh Symposium of the Federation of Acoustic Societies of Europe, vol. 3*, pages 827–833, Edinburgh.

Morten H. Christiansen, Joseph Allen, and Mark S. Seidenberg. 1998. Learning to Segment Speech Using Multiple Cues: A Connectionist Model. *Language and Cognitive Processes*, 13(2/3):221–269.

C. M. Connine, D. Titone, and J. Wang. 1993. Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19:81–94.

Anne Cutler. 1990. Exploiting prosodic probabilities in speech segmentation. In G. A. Altmann, editor, *Cognitive models of speech processing: Psycholinguistic and computational perspectives*, pages 105–121. MIT Press, Cambridge, MA.

Robert Daland and Janet B. Pierrehumbert. 2011. Learning diphone-based segmentation. *Cognitive Science*, 35(1):119–155.

Laura C. Dilley and Mark Pitt. 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11):1664–1670.

Markus Dreyer, Jason R. Smith, and Jason Eisner. 2008. Latent-variable modeling of string transductions with finite-state methods. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 1080–1089, Stroudsburg, PA, USA. Association for Computational Linguistics.

Emmanuel Dupoux, Guillaume Beraud-Sudreau, and Shigeki Sagayama. 2011. Templatic features for modeling phoneme acquisition. In *Proceedings of the 33rd Annual Cognitive Science Society*.

Micha Elsner, Sharon Goldwater, and Jacob Eisenstein. 2012. Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 184–193, Jeju Island, Korea, July. Association for Computational Linguistics.

Naomi Feldman, Thomas Griffiths, and James Morgan. 2009. Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

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

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

Margaret M. Fleck. 2008. Lexicalized phonotactic word segmentation. In *Proceedings of ACL-08: HLT*, pages 130–138, Columbus, Ohio, June. Association for Computational Linguistics.

Michael C. Frank, Sharon Goldwater, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2010. Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–125.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Rebecca Gómez and Jessica Maye. 2005. The developmental trajectory of nonadjacent dependency learning. *Infancy*, 7:183–206.

Bruce Hayes and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Marieke van Heugten and Elizabeth K. Johnson. in press. Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*.

Derek M. Houston and Peter W. Jusczyk. 2000. The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26:1570–1582.

Jonathan Huggins and Frank Wood. 2013. Infinite structured hidden semi-Markov models. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, to appear, September.

Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard Rose, Mike Seltzer, Pascal Clark, Ian McGraw, Balakrishnan Varadarajan, Erin Bennett, Benjamin Borschinger, Justin Chiu, Ewan Dunbar, Abdellah Fourtassi, David Harwath, Chia-ying Lee, Keith Levin, Atta Norouzian, Vijay Peddinti, Rachael Richardson, Thomas Schatz, and Samuel Thomas. 2013. A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and early language acquisition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Mark Johnson. 2008. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of ACL-08: HLT*, pages 398–406, Columbus, Ohio, June. Association for Computational Linguistics.

Peter W. Jusczyk and Richard N. Aslin. 1995. Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29:1–23.

Peter W. Jusczyk, Derek M. Houston, and Mary Newsome. 1999. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, 39:159–207.

Dahee Kim, Joseph D.W. Stephens, and Mark A. Pitt. 2012. How does context play a part in splitting words apart? Production and perception of word boundaries in casual speech. *Journal of Memory and Language*, 66(4):509 – 529.

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

Leigh Lisker and Arthur S. Abramson. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20:384–422.

Andrew Martin, Sharon Peperkamp, and Emmanuel Dupoux. 2013. Learning phonemes with a proto-lexicon. *Cognitive Science*, 37:103–124.

Sven L. Mattys and Peter W. Jusczyk. 2001. Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, 27(3):644–655+.

Jessica Maye, Janet F. Werker, and LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–11.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 100–108, Suntec, Singapore, August. Association for Computational Linguistics.

Mehryar Mohri, 2004. *Weighted Finite-State Transducer Algorithms: An Overview*, chapter 29, pages 551–564. Physica-Verlag.

Thierry Nazzi, Laura C. Dilley, Ann Marie Jusczyk, Stefanie Shattuck-Hufnagel, and Peter W. Jusczyk. 2005. English-learning infants' segmentation of verbs from fluent speech. *Language and Speech*, 48(3):279–298+.

Graham Neubig, Masato Mimura, Shinsuke Mori, and Tatsuya Kawahara. 2010. Learning a language model from continuous speech. In *11th Annual Conference of the International Speech Communication Association (InterSpeech 2010)*, pages 1053–1056, Makuhari, Japan, 9.

Sharon Peperkamp, Rozenn Le Calvez, Jean-Pierre Nadal, and Emmanuel Dupoux. 2006. The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41.

Ann M. Peters. 1983. *The Units of Language Acquisition*. Cambridge Monographs and Texts in Applied Psycholinguistics. Cambridge University Press.

Gordon E. Peterson and Harold L. Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2):175–184.

Mark A. Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2007. Buckeye corpus of conversational speech (2nd release).

Kim Plunkett. 2005. Learning how to be flexible with words. *Attention and Performance*, XXI:233–248.

Anton Rytting. 2007. *Preserving Subsegmental Variation in Modeling Word Segmentation (Or, the Raising of Baby Mondegreen)*. Ph.D. thesis, The Ohio State University.

Amanda Seidl and Elizabeth Johnson. 2006. Infant word segmentation revisited: Edge alignment facilitates target extraction. *Developmental Science*, 9:565–573.

Amanda Seidl and Elizabeth Johnson. 2008. Perceptual factors influence infants' extraction of onsetless words from continuous speech. *Journal of Child Language*, 34.

Leher Singh, James Morgan, and Katherine White. 2004. Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, 51:173–189.

Daniel Swingley. 2005. Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, 50:86–132.

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

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Michael Tomasello. 2000. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4(4):156 – 163.

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

Jurgen Van Gael, Yunus Saatci, Yee Whye Teh, and Zoubin Ghahramani. 2008. Beam sampling for the infinite Hidden Markov model. In *Proceedings of the 25th International Conference on Machine learning*, ICML '08, pages 1088–1095, New York, NY, USA. ACM.

Balakrishnan Varadarajan, Sanjeev Khudanpur, and Emmanuel Dupoux. 2008. Unsupervised learning of acoustic sub-word units. In *Proceedings of the Association for Computational Linguistics: Short Papers*, pages 165–168.

Anand Venkataraman. 2001. A statistical model for word discovery in transcribed speech. *Computational Linguistics*, 27(3):351–372.

Janet F. Werker and Richard C. Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49 – 63.

Katherine S. White and Richard N. Aslin. 2011. Adaptation to novel accents by toddlers. *Developmental Science*, 14(2):372–384.