Unsupervised Lexicon Discovery from Acoustic Input

Chia-ying Lee[1]**, Timothy J. O'Donnell**[2]**, and James Glass**[1]
[1]Computer Science and Artificial Intelligence Laboratory
[2]Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Abstract

We present a model of *unsupervised phonological lexicon discovery*—the problem of simultaneously learning phoneme-like and word-like units from acoustic input. Our model builds on earlier models of unsupervised phone-like unit discovery from acoustic data (Lee and Glass, 2012), and unsupervised symbolic lexicon discovery using the Adaptor Grammar framework (Johnson et al., 2006), integrating these earlier approaches using a probabilistic model of phonological variation. We show that the model is competitive with state-of-the-art spoken term discovery systems, and present analyses exploring the model's behavior and the kinds of linguistic structures it learns.

1 Introduction

One of the most basic problems of language acquisition is accounting for how children learn the inventory of word forms from speech—*phonological lexicon discovery*. In learning a language, children face a number of challenging, mutually interdependent inference problems. Words are represented in terms of *phonemes*, the basic phonological units of a language. However, phoneme inventories vary from language to language, and the underlying phonemes which make up individual words often have variable acoustic realizations due to systematic phonetic and phonological variation, dialect differences, speech style, environmental noise, and other factors. To learn the phonological form of words in their language children must determine the phoneme inventory of their language, identify which parts of the acoustic signal correspond to which phonemes—while discounting surface variation in the realization of individual units—and infer which sequences of phonemes correspond to which words (amongst other challenges).

Understanding the solution to this complex joint-learning problem is not only of fundamental scientific interest, but also has important applications in Spoken Language Processing (SLP). Even setting aside additional grammatical and semantic information available to child learners, there is still a sharp contrast between the type of phonological learning done by humans and current SLP methods. Tasks that involve recognizing words from acoustic input—such as automatic speech recognition and spoken term discovery—only tackle parts of the overall problem, and typically rely on linguistic resources such as phoneme inventories, pronunciation dictionaries, and annotated speech data. Such resources are unavailable for many languages, and expensive to create. Thus, a model that can jointly learn the sound patterns and the lexicon of a language would open up the possibility of automatically developing SLP capabilities for any language.

In this paper, we present a first step towards an unsupervised model of phonological lexicon discovery that is able to jointly learn, from unannotated speech, an underlying phoneme-like inventory, the pattern of surface realizations of those units, and a set of lexical units for a language. Our model builds on earlier work addressing the unsupervised discovery of phone-like units from acoustic data—in particular the Dirichlet Process Hidden Markov Model (DPHMM) of Lee and Glass (2012)—and the un-

Transactions of the Association for Computational Linguistics, vol. 3, pp. 389–403, 2015. Action Editor: Eric Fosler-Lussier.
Submission batch: 11/2014; Revision batch 2/2015; Published 7/2015.

supervised learning of lexicons from unsegmented symbolic sequences using the Adaptor Grammar (AG) framework (Johnson et al., 2006; Johnson and Goldwater, 2009a). We integrate these models with a component modeling variability in the surface realization of phoneme-like units within lexical units.

In the next section, we give an overview of related work. Following this, we present our model and inference algorithms. We then turn to preliminary evaluations of the model's performance, showing that the model is competitive with state-of-the-art single-speaker spoken term discovery systems, and providing several analyses which examine the kinds of structures learned by the model. We also suggest that the ability of the system to successfully unify multiple acoustic sequences into single lexical items relies on the phonological-variability (noisy channel) component of the model—demonstrating the importance of modeling symbolic variation in phonological units. We provide preliminary evidence that simultaneously learning sound and lexical structure leads to *synergistic* interactions (Johnson, 2008b)—the various components of the model mutually constrain one another such that the linguistic structures learned by each are more accurate than if they had been learned independently.

2 Related Work

Previous models of lexical unit discovery have primarily fallen into two classes: models of *spoken term discovery* and models of *word segmentation*. Both kinds of models have sought to identify lexical items from input without direct supervision, but have simplified the joint learning problem discussed in the introduction in different ways.

Spoken Term Discovery Spoken term discovery is the problem of using unsupervised pattern discovery methods to find previously unknown keywords in speech. Most models in this literature have typically made use of a two-stage procedure: First, subsequences of the input that are similar in an acoustic feature space are identified, and, then clustered to discover categories corresponding to lexical items (Park and Glass, 2008; Zhang and Glass, 2009; Zhang et al., 2012; Jansen et al., 2010; Aimetti, 2009; McInnes and Goldwater, 2011). This problem was first examined by Park and Glass (2008) who used Dynamic Time Warping to identify similar acoustic sequences across utterances. The input sequences discovered by this method were then treated as nodes in a similarity-weighted graph, and graph clustering algorithms were applied to produce a number of densely connected groups of acoustic sequences, corresponding to lexical items. Building on this work, Zhang and Glass (2009) and Zhang et al. (2012) proposed robust features that allowed lexical units to be discovered from spoken documents generated by different speakers. Jansen et al. (2010) present a similar framework for finding repeated acoustic patterns, based on line-segment detection in dotplots. Other variants of this approach include McInnes and Goldwater (2011) who compute similarity incrementally, and Aimetti (2009) who integrates a simplified, symbolic representation of visual information associated with each utterance.

Word Segmentation In contrast to spoken term discovery, models of word (or morpheme) segmentation start from unsegmented strings of symbols and attempt to identify subsequences corresponding to lexical items. The problem has been the focus of many years of intense research, and there are a large variety of proposals in the literature (Harris, 1955; Saffran et al., 1996a; Harris, 1955; Olivier, 1968; Saffran et al., 1996b; Brent, 1999b; Frank et al., 2010; Frank et al., 2013). Of particular interest here are models which treat segmentation as a secondary consequence of discovering a compact lexicon which explains the distribution of phoneme sequences in the input (Cartwright and Brent, 1994; Brent, 1999a; Goldsmith, 2001; Argamon et al., 2004; Goldsmith, 2006; Creutz and Lagus, 2007; Goldwater et al., 2009; Mochihashi et al., 2009; Elsner et al., 2013; Neubig et al., 2012; Heymann et al., 2013; De Marcken, 1996c; De Marcken, 1996a). Recently, a number of such models have been introduced which make use of Bayesian nonparametric distributions such as the Dirichlet Process (Ferguson, 1973) or its two-parameter generalization, the Pitman-Yor Process (Pitman, 1992), to define a prior which favors smaller lexicons with more reusable lexical items. The first such models were proposed in Goldwater (2006) and, subsequently, have been extended in a number of ways (Goldwater et al., 2009; Neubig et al., 2012; Heymann et al., 2013;

Mochihashi et al., 2009; Elsner et al., 2013; Johnson et al., 2006; Johnson and Demuth, 2010; Johnson and Goldwater, 2009b; Johnson, 2008a; Johnson, 2008b).

One important lesson that has emerged from this literature is that models which jointly represent multiple levels of linguistic structure often benefit from *synergistic* interactions (Johnson, 2008b) where different levels of linguistic structure provide mutual constraints on one another which can be exploited simultaneously (Goldwater et al., 2009; Johnson et al., 2006; Johnson and Demuth, 2010; Johnson and Goldwater, 2009b; Johnson, 2008a; Börschinger and Johnson, 2014; Johnson, 2008b). For example, Elsner et al. (2013) show that explicitly modeling symbolic variation in phoneme realization improves lexical learning—we use a similar idea in this paper.

An important tool for studying such synergies has been the Adaptor Grammars framework of Johnson et al. (2006). Adaptor Grammars are a generalization of Probabilistic Context-free Grammars (PCFGs) which allow the lexical storage of complete subtrees. Using Adaptor Grammars, it is possible to learn lexica which contain stored units at multiple levels of abstraction (e.g., phonemes, onsets, codas, syllables, morphemes, words, and multiword collocations). A series of studies using the framework has shown that including such additional structure can markedly improve lexicon discovery (Johnson et al., 2006; Johnson and Demuth, 2010; Johnson and Goldwater, 2009b; Johnson, 2008a; Börschinger and Johnson, 2014; Johnson, 2008b).

Unsupervised Lexicon Discovery In contrast to models of spoken term discovery and word segmentation, our model addresses the problem of jointly inferring phonological and lexical structure directly from acoustic input. Spoken term discovery systems only attempt to detect keywords, finding lexical items that are isolated and scattered throughout the input data. They do not learn any intermediate levels of linguistic structure between the acoustic input and discovered lexical items. In constrast, our model attempts to find a complete segmentation of the input data into units at multiple levels of abstraction (e.g., phonemes, syllables, words, etc.). Unlike word segmentation models, our model works directly from speech input, integrating unsupervised acoustic modeling with an an approach to symbolic lexicon discovery based on adaptor grammars.

Although some earlier systems have examined various parts of the joint learning problem (Bacchiani and Ostendorf, 1999; De Marcken, 1996b), to our knowledge, the only other system which addresses the entire problem is that of Chung et al. (2013). There are two main differences between the approaches. First, in Chung et al. (2013), word-like units are defined as unique sequences of sub-word-like units, so that any variability in the realization of word-parts must be accounted for by the acoustic models. In contrast, we explicitly model phonetic variability at the symbolic level, allowing our system to learn low-level units which tightly predict the acoustic realization of phonemes in particular contexts, while still ignoring this variability when it is irrelevant to distinguishing lexical items. Second, while Chung et al. (2013) employ a fixed, two-level representation of linguistic structure, our use of adaptor grammars to model symbolic lexicon discovery means that we can easily and flexibly vary our assumptions about the hierarchical makeup of utterances and lexical items. In this paper, we employ a simple adaptor grammar with three-levels of hierarchical constituency (word-like, sub-word-like, and phone-like units) and with right-branching structure; future work could explore more articulated representations along the lines of Johnson (2008b).

3 Model

3.1 Problem Formulation and Model Overview

Given a corpus of spoken utterances, our model aims to jointly infer the phone-like, sub-lexical, and lexical units in each spoken utterance. To discover these hierarchical linguistic structures directly from acoustic signals, we divide the problem into three sub-tasks: 1) phone-like unit discovery, 2) variability modeling, and 3) sub-lexical and lexical unit learning. Each of the sub-tasks corresponds to certain latent structures embedded in the speech data that our model must identify. Here we briefly discuss the three sub-tasks as well as the latent variables associated with each, and provide an overview on the proposed model for each sub-task.

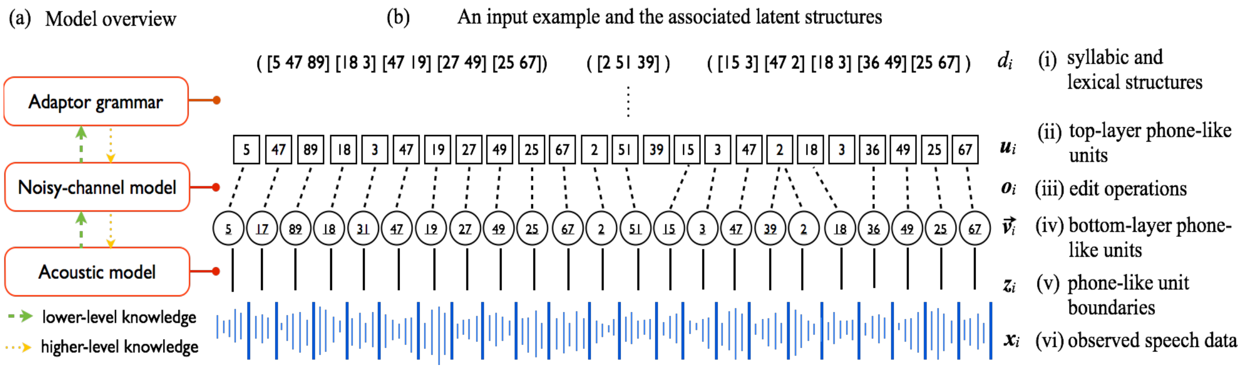
(a) Model overview
Adaptor grammar
Noisy-channel model
Acoustic model
lower-level knowledge
higher-level knowledge
(b) An input example and the associated latent structures
([5 47 89] [18 3] [47 19] [27 49] [25 67]) ([2 51 39]) ([15 3] [47 2] [18 3] [36 49] [25 67])
5 47 89 18 3 47 19 27 49 25 67 2 51 39 15 3 47 2 18 3 36 49 25 67
5 17 89 18 31 47 19 27 49 25 67 2 51 15 3 47 39 2 18 36 49 25 67
d_i (i) syllabic and lexical structures
$\boldsymbol{u}_i$ (ii) top-layer phone-like units
$\boldsymbol{o}_i$ (iii) edit operations
$\vec{\boldsymbol{v}}_i$ (iv) bottom-layer phone-like units
$\boldsymbol{z}_i$ (v) phone-like unit boundaries
$\boldsymbol{x}_i$ (vi) observed speech data

Figure 1: (a) An overview of the proposed model for inducing hierarchical linguistic structures directly from acoustic signals. As indicated in the graph, the model leverages partial knowledge learned from each level to drive discovery in the others. (b) An illustration of an input example, $\boldsymbol{x}_i$, and the associated latent structures in the acoustic signals $d_i, \boldsymbol{u}_i, \boldsymbol{o}_i, \vec{\boldsymbol{v}}_i, \boldsymbol{z}_i$. These latent structures can each be discovered by one of the three components of the model as specified by the red horizontal bars between (a) and (b).

Phone-like unit discovery For this sub-task, the model converts the speech input $\boldsymbol{x}_i$ into a sequence of Phone-Like Units (PLUs), $\vec{\boldsymbol{v}}_i$, which implicitly determines the phone segmentation, $\boldsymbol{z}_i$, in the speech data as indicated in (iv)-(vi) of Fig 1-(b). We use $\boldsymbol{x}_i = \{x_{i,t} | x_{i,t} \in \mathbb{R}^{39}, 1 \leq t \leq T_i\}$ to denote the series of Mel-Frequency Cepstral Coefficients (MFCCs) representing the i^{th} utterance (Davis and Mermelstein, 1980), where T_i stands for the total number of feature frames in utterance i. Each $\boldsymbol{x}_i$ contains 13-dimensional MFCCs and their first- and second-order time derivatives at a 10 ms frame rate. Each $\boldsymbol{x}_i$ is also associated with a binary variable $z_{i,t}$, indicating whether a PLU boundary exists between $x_{i,t}$ and $x_{i,t+1}$. The feature vectors with $z_{i,t} = 1$ are highlighted by the dark blue bars in Fig. 1-(vi), which correspond to segment boundaries.

Each speech segment is labelled with a PLU id $v_{i,j,k} \in \mathbb{L}$, in which $\mathbb{L}$ is a set of integers that represent the PLU inventory embedded in the speech corpus. We denote the sequence of PLU ids associated with utterance i using $\vec{\boldsymbol{v}}_i$ as shown in Fig. 1-(iv), where $\vec{\boldsymbol{v}}_i = \{\boldsymbol{v}_{i,j} | 1 \leq j \leq J_i\}$ and $\boldsymbol{v}_{i,j} = \{v_{i,j,k} | v_{i,j,k} \in \mathbb{L}, 1 \leq k \leq |\boldsymbol{v}_{i,j}|\}$. The variable J_i is defined in the discussion of the second sub-task. As depicted in Fig. 1-(a), we construct an acoustic model to approach this sub-problem. The acoustic model is composed of a set of Hidden Markov Models (HMMs), $\boldsymbol{\pi}$, that are used to infer and model the PLU inventory from the given data.

Phonological variability modeling In conversational speech, the phonetic realization of a word can easily vary because of phonological and phonetic context, stress pattern, etc. Without a mechanism that can map these speech production variations into a unique representation, any model that induces linguistic structures based on phonetic input would fail to recognize these pronunciations as instances of the same word type. We exploit a noisy-channel model to address this problem and design three edit operations for the noisy-channel model: substitute, split, and delete. Each of the operations takes a PLU as an input and is denoted as $\text{sub}(u)$, $\text{split}(u)$, and $\text{del}(u)$ respectively. We assume that for every inferred sequence of PLUs $\vec{\boldsymbol{v}}_i$ in Fig. 1-(b)-(iv), there is a corresponding series of PLUs, $\boldsymbol{u}_i = \{u_{i,j} | 1 \leq j \leq J_i\}$, in which the pronunciations for any repeated word in $\vec{\boldsymbol{v}}_i$ are identical. The variable J_i indicates the length of $\boldsymbol{u}_i$. By passing each $u_{i,j}$ through the noisy-channel model, which stochastically chooses an edit operation $o_{i,j}$ for $u_{i,j}$, we obtain the noisy phonetic realization $\boldsymbol{v}_{i,j}$.[1] The relationship among $\boldsymbol{u}_i$, $\boldsymbol{o}_i$, and $\vec{\boldsymbol{v}}_i$ is shown in (ii)-(iv) of Fig. 1-(b). For notation simplicity, we let $o_{i,j}$ encode $u_{i,j}$ and $\boldsymbol{v}_{i,j}$, which means we can read $u_{i,j}$ and $\boldsymbol{v}_{i,j}$ directly from $o_{i,j}$. We refer to the units that are input to the noisy-channel model $\boldsymbol{u}_i$ as "top-layer" PLUs and the units that are output from the noisy-channel model $\vec{\boldsymbol{v}}_i$ as "bottom-layer" PLUs.

[1] We denote $\boldsymbol{v}_{i,j}$ as a vector since if a split operation is chosen for $u_{i,j}$, the noisy-channel model will output two PLUs.

Sub-word-like and word-like unit learning With the standardized phone-like representation $\boldsymbol{u}_i$, higher-level linguistic structures can be inferred for each spoken utterance. We employ Adaptor Grammars (AGs) (Johnson et al., 2006) to achieve this goal, and use d_i to denote the parse tree that encodes the hierarchical linguistic structures in Fig. 1-(b)-(i). We bracket sub-word-like (e.g., syllable) and word-like units using $[\cdot]$ and $(\cdot)$, respectively.

In summary, our model integrates adaptor grammars with a noisy-channel model of phonetic variability and an acoustic model to discover hierarchical linguistic structures directly from acoustic signals. Even though we have discussed these sub-tasks in a bottom-up manner, our model provides a joint learning framework, allowing knowledge learned from one sub-task to drive discovery in the others. We now a review the formalization of adaptor grammars and define the noisy-channel and acoustic components of our model. We conclude this section by presenting the generative process implied by the three components of our model.

3.2 Adaptor Grammars

Adaptor grammars are a non-parametric Bayesian extension of Probabilistic Context-Free Grammars (PCFGs). A PCFG can be defined as a quintuple $(N, T, R, S, \{\vec{\theta}^q\}_{q \in N})$, which consists of disjoint finite sets of non-terminal symbols N and terminal symbols T, a finite set of production rules $R \subseteq \{N \to (N \cup T)^*\}$, a start symbol $S \in N$, and vectors of probabilistic distributions $\{\vec{\theta}^q\}_{q \in N}$. Each $\vec{\theta}^q$ contains the probabilities associated with the rules that have the non-terminal q on their left-hand side. We use θ_r to indicate the probability of rule $r \in R$. We adopt a Bayesian approach and impose a Dirichlet prior on each $\vec{\theta}^q \sim Dir(\vec{\alpha}^q)$.

Let t denote a *complete derivation*, which represents either a tree that expands from a non-terminal q to its leaves, which contain only terminal symbols, or a tree that is composed of a single terminal symbol. We define $\mathrm{root}(t)$ as a function that returns the root node of t and denote the k immediate subtrees of the root node as $\hat{t}_1, \cdots, \hat{t}_k$. The probability distribution over $\mathcal{T}^q$, the set of trees that have $q \in N \cup T$ as the root, is recursively defined as follows.

$$G^q_{\text{pcfg}}(t) = \begin{cases} \sum_{r \in R^q} \theta_r \prod_{i=1}^{k} G_{\text{pcfg}}^{\mathrm{root}(\hat{t}_i)}(\hat{t}_i) & \mathrm{root}(t) = q \in N \\ 1 & \mathrm{root}(t) = q \in T \end{cases}$$

An adaptor grammar is a sextuple $(N, T, R, S, \{\vec{\theta}^q\}_{q \in N}, \{Y^q\}_{q \in N})$, in which $(N, T, R, S, \{\vec{\theta}^q\}_{q \in N})$ is a PCFG, and $\{Y^q\}_{q \in N}$ is a set of *adaptors* for the non-terminals. An adaptor Y^q is a function that maps a base distribution over $\mathcal{T}^q$ to *a distribution* over distributions over $\mathcal{T}^q$. The distribution $G^q_{\text{ag}}(t)$ for $q \in N$ of an AG is a sample from this distribution over distributions. Specifically,

$$G^q_{\text{ag}}(t) \sim Y^q(H^q(t))$$
$$H^q(t) = \sum\nolimits_{r \in R^q} \theta_r \prod\nolimits_{i=1}^{k} G_{\text{ag}}^{\mathrm{root}(\hat{t}_i)}(\hat{t}_i),$$

where $H^q(t)$ denotes the base distribution over $\mathcal{T}^q$. In this paper, following Johnson et al. (2006), we use adaptors that are based on Pitman-Yor processes (Pitman and Yor, 1997). For terminal symbols $q \in T$, we define $G^q_{\text{ag}}(t) = 1$, which is a distribution that puts all its probability mass on the single-node tree labelled q. Conceptually, AGs can be regarded as PCFGs with memories that cache the complete derivations of adapted non-terminals, allowing the AG to choose to either reuse the cached trees or select a production an underlying rule in R to expand each non-terminal. For a more detailed description of AGs and their connection to PCFGs, we refer readers to Johnson et al. (2006) and Chapter 3 of O'Donnell (2015).

To discover the latent hierarchical linguistic structures in spoken utterances, we employ the following AG to parse each spoken utterance, where we adopt the notations of Johnson and Goldwater (2009b) and use underlines to indicate adapted non-terminals and employ $^+$ to abbreviate right-branching recursive rules for non-terminals. The last rule shows that the terminals of this AG are the PLU ids, which are represented as $\boldsymbol{u}_i$ and depicted as the units in the squares of Fig. 1-(b)-(ii).

$$\text{Sentence} \to \underline{\text{Word}}^+$$
$$\underline{\text{Word}} \to \underline{\text{Sub-word}}^+$$
$$\underline{\text{Sub-word}} \to \text{Phone}^+$$
$$\text{Phone} \to l \quad \text{for } l \in \mathbb{L}$$

Note that the grammar above only makes use of right-branching rules and therefore could be simulated using finite-state infrastructure, rather than the more complex context-free machinery implicit in the adaptor grammars framework. We nevertheless make use of the formalism for two reasons. First, on a theoretical level it provides a uniform framework for expressing many different assumptions about the symbolic component of segmentation models (Goldwater et al., 2009; Johnson, 2008b; Börschinger and Johnson, 2014; Johnson, 2008a; Johnson and Goldwater, 2009b; Johnson and Demuth, 2010). Using adaptor grammars to formalize the symbolic component our our model thus allows direct comparisons to this literature as well as transparent extensions following earlier work. Second, on a practical level, using the framework allowed us to make use of Mark Johnson's efficient implementation of the core adaptor grammar sampling loop, significantly reducing model development time.

3.3 Noisy-channel Model

We formulate the noisy-channel model as a PCFG and encode the substitute, split, and delete operations as grammar rules. In particular, for $l \in \mathbb{L}$,

$$\begin{aligned} & l \rightarrow l' && \text{for } l' \in \mathbb{L} \\ & l \rightarrow 1'_1 l'_2 && \text{for } l'_1, l'_2 \in \mathbb{L} \\ & l \rightarrow \epsilon && \end{aligned} \tag{1}$$

where $l \in \mathbb{L}$ are the start symbols as well as the non-terminals of the PCFG. The terminals of this PCFG are $l' \in \mathbb{L}$, which correspond to bottom-layer PLUs $\vec{\boldsymbol{v}}_i$ that are depicted as units in circles in Fig. 1-(b)-(iv). Note that $\{l\}$ and $\{l'\}$ are drawn from the same inventory of PLUs, and the notation is meant to signal that $\{l'\}$ are the terminal symbols of this grammar. The three sets of rules respectively map to the $\mathrm{sub}(\cdot)$, $\mathrm{split}(\cdot)$, and $\mathrm{del}(\cdot)$ operations; thus, the probability of each edit operation is automatically captured by the corresponding rule probability. Note that to simultaneously infer a phonetic inventory of an unknown size and model phonetic variation, we can use the infinite PCFG (Liang et al., 2007) to formulate the noisy-channel model. However, for computational efficiency, in our experiments, we infer the size of the PLU inventory before training the full model, and impose a Dirichlet prior on the rule probability distribution associated with each non-terminal l. We explain how inventory size is determined in Sec. 5.2.

3.4 Acoustic Model

Finally, we assign each discovered PLU $l \in \mathbb{L}$ to an HMM, π_l, which is used to model the speech realization of each phonetic unit in the feature space. In particular, to capture the temporal dynamics of the features associated with a PLU, each HMM contains three emission states, which roughly correspond to the beginning, middle, and end of a phonetic unit (Jelinek, 1976). We model the emission distribution of each state by using 39-dimensional diagonal Gaussian Mixture Models (GMMs). The prior distributions embedded in the HMMs are the same as those described in (Lee and Glass, 2012).

3.5 Generative Process of the Proposed Model

With the adaptor grammar, the noisy-channel model, and the acoustic model defined, we summarize the generative process implied by our model as follows. For the i^{th} utterance in the corpus, our model

1. Generates a parse tree d_i from $G_{\text{ag}}^{\text{Sentence}}(\cdot)$.
2. For each leaf node $u_{i,j}$ of d_i, samples an edit rule $o_{i,j}$ from $\vec{\theta}^{u_{i,j}}$ to convert $u_{i,j}$ to $\boldsymbol{v}_{i,j}$.
3. For $v_{i,j,k} \in \boldsymbol{v}_{i,j}$, $1 \leq k \leq |\boldsymbol{v}_{i,j}|$, generates the speech features using $\pi_{v_{i,j,k}}$, which deterministically sets the value of $z_{i,t}$.

Thus, the latent variables our model defines for each utterance are: d_i, $\boldsymbol{u}_i$, $\boldsymbol{o}_i$, $\vec{\boldsymbol{v}}_i$, $\boldsymbol{z}_i$, $\boldsymbol{\pi}$, $\{\vec{\theta}^q\}_{q \in N_{\text{ag}} \cup N_{\text{noisy-channel}}}$. In the next section, we derive inference methods for all the latent variables except for $\{\vec{\theta}^q\}_{q \in N_{\text{ag}} \cup N_{\text{noisy-channel}}}$, which we integrate out during the inference process.

4 Inference

We exploit Markov chain Monte Carlo algorithms to generate samples from the posterior distribution over the latent variables. In particular, we construct three Markov chain kernels: 1) jointly sampling d_i, $\boldsymbol{o}_i$, $\boldsymbol{u}_i$, 2) generating new samples for $\vec{\boldsymbol{v}}_i$, $\boldsymbol{z}_i$, and 3) updating $\boldsymbol{\pi}$. Here, we give an overview of each of the sampling moves.

4.1 Sampling d_i, $\boldsymbol{o}_i$, and implicitly $\boldsymbol{u}_i$

We employ the Metropolis-Hastings (MH) algorithm (Chib and Greenberg, 1995) to generate samples for d_i and $\boldsymbol{o}_i$, which implicitly determines $\boldsymbol{u}_i$.

Given $d_{-i}, \boldsymbol{o}_{-i}$, the current parses and the current edit operations associated with all the sentences in the corpus except the i^{th} utterance, we can construct a proposal distribution for d'_i and $\boldsymbol{o}'_i$[2] by using the approximating PCFG described in Johnson et al. (2006) and the approximated probability of $o'_{i,j}$ in $\boldsymbol{o}'_i$, which is defined in Eq. 2.

$$p(o'_{i,j}|\boldsymbol{o}_{-i}; \{\vec{\alpha}\}^q) \approx \frac{C_{-i}(u'_{i,j} \rightarrow \boldsymbol{v}'_{i,j}) + \vec{\alpha}^{u'_{i,j}}_{u'_{i,j} \rightarrow \boldsymbol{v}'_{i,j}}}{C_{-i}(u'_{i,j}) + \sum_{r \in R^{u'_{i,j}}} \vec{\alpha}^{u'_{i,j}}_r} \tag{2}$$

where $q \in N_{\text{noisy-channel}}$, and $C_{-i}(w)$ denotes the number of times that w is used in the analyses for the corpus, excluding the i^{th} utterance, in which w can be any countable entity such as a rule or a symbol.

More specifically, we combine the PCFG that approximates the adaptor grammar with the noisy-channel PCFG whose rules are weighted as in Eq. 2 to form a new PCFG G'. The new PCFG G' is thus a grammar that can be used to parse the terminals $\vec{\boldsymbol{v}}_i$ and generate derivations that are rooted at the start symbol of the AG. Therefore, we transform the task of sampling d'_i and $\boldsymbol{o}'_i$ to the task of generating a parse for $\vec{\boldsymbol{v}}_i$ using G', which can be efficiently solved by using an variant of the Inside algorithm for PCFGs (Lari and Young, 1990; Johnson et al., 2007; Goodman, 1998; Finkel et al., 2006).

4.2 Sampling $\vec{\boldsymbol{v}}_i$ and $\boldsymbol{z}_i$

Given the top-layer PLUs $\boldsymbol{u}_i$ and the speech data $\boldsymbol{x}_i$, sampling the boundary variables $\boldsymbol{z}_i$ and the bottom-layer PLUs $\vec{\boldsymbol{v}}_i$ is equivalent to sampling an alignment between $\boldsymbol{u}_i$ and $\boldsymbol{x}_i$. Therefore, we use the probabilities defined in Eq. 2 and employ the backward message-passing and forward-sampling algorithm described in Lee et al. (2013), designed for aligning a letter sequence and speech signals, to propose samples for $\vec{\boldsymbol{v}}_i$ and $\boldsymbol{z}_i$. The proposals are then accepted by using the standard MH criterion.

4.3 Sampling $\boldsymbol{\pi}$

Given $\boldsymbol{z}_i$ and $\vec{\boldsymbol{v}}_i$ of each utterance in the corpus, generating new samples for the parameters of each HMM π_l for $l \in \mathbb{L}$ is straightforward. For each PLU l, we gather all speech segments that are mapped to a bottom-layer PLU $v_{i,j,k} = l$. For every segment in this set, we use π_l to block-sample the state id and the GMM mixture id for each feature vector. From the state and mixture assignments, we can collect the counts that are needed to update the priors for the transition probability and the emission distribution of each state in π_l. New samples for the parameters of π_l can thus be yielded from the updated priors.

Lecture topic	Duration
Economics	75 mins
Speech processing	85 mins
Clustering	78 mins
Speaker adaptation	74 mins
Physics	51 mins
Linear algebra	47 mins

Table 1: A brief summary of the six lectures used for the experiments reported in Section 6.

5 Experimental Setup

5.1 Dataset

To the best of our knowledge, there are no standard corpora for evaluating models of unsupervised lexicon discovery. In this paper, we perform experiments on the six lecture recordings used in (Park and Glass, 2008; Zhang and Glass, 2009), a part of the MIT Lecture corpus (Glass et al., 2004). A brief summary of the six lectures is listed in Table 1.

5.2 Systems

Full system We constructed two systems based on the model described in Sec. 3. These systems, FullDP and Full50, differ only in the size of the PLU inventory (K). For FullDP, we set the value of K to be the number of PLUs discovered for each lecture by the DPHMM framework presented in (Lee and Glass, 2012). These numbers were: Economics, 99; Speech Processing, 111; Clustering, 91; Speaker Adaptation, 83; Physics, 90; and Algebra, 79. For Full50, we used a fixed number of PLUs, $K = 50$.

The acoustic component of the FullDP system was initialized by using the output of the DPHMM model for each lecture. Specifically, we made use of the HMMs, the phone boundaries, and the PLU that the DPHMM model found as the initial values for $\boldsymbol{\pi}$, $\boldsymbol{z}_i$, and $\vec{\boldsymbol{v}}_i$ of the FullDP system. After initialization, the training of FullDP proceeds by following

[2] We use d_i and d'_i to denote the current and the proposed parses. The same relationship is also defined for $\boldsymbol{o}_i$ and $\boldsymbol{o}'_i$.

the three sampling moves described in Sec. 4. Similarly, we employ a Hierarchical HMM (HHMM), which is presented in detail in (Lake et al., 2014), to find the initial values of $\boldsymbol{\pi}$, $\boldsymbol{z}_i$, and $\vec{\boldsymbol{v}}_i$ for the Full50 system. The adaptor grammar component of all systems was initialized following the "batch initialization" method described in Johnson and Goldwater (2009b) which independently samples an AG analysis for each utterance. The lesioned systems that are described in the rest of this section were also initialized in the same manner.

To reduce the inference load on the HMM, we exploit acoustic cues in the feature space to constrain phonetic boundaries to occur at a subset of all possible locations (Lee and Glass, 2012). We follow the pre-segmentation method described in (Glass, 2003) to achieve the goal. Empirically, this boundary elimination heuristic reduces the computational complexity of the inference algorithm by roughly an order of magnitude on clean speech corpora.

No acoustic model We remove the acoustic model from FullDP and Full50 to obtain the -AM systems. Since the -AM systems do not have an acoustic model, they cannot resegment or relabel the data, which implies that there is no learning of phonetic units in the -AM systems, making them similar to symbolic segmentation models that include a noisy channel component (Elsner et al., 2013). By comparing a -AM system to its full counterpart, we can investigate the synergies between phonetic and lexical unit acquisition in the full model.

No noisy-channel To evaluate the importance of modeling phonetic variability, we remove the noisy-channel model from the -AM systems to form -NC systems. A -NC system can be regarded as a pipeline, whereby utterance phone sequences are discovered first, and latent higher-level linguistic structures are learned in the second step and thus is similar to models such as that of Jansen et al. (2010).

6 Results and Analyses

Training convergence Fig. 2 shows the negative log posterior probability of the sampled parses $\boldsymbol{d}$ and edit operations $\boldsymbol{o}$ for each lecture as a function of iteration generated by the FullDP system. Given that each lecture consists of roughly only one hour of speech data, we can see that the model converges fairly quickly within a couple hundreds of iterations. In the this section, we report the performance of each system using the corresponding sample from the 200^{th} iteration.

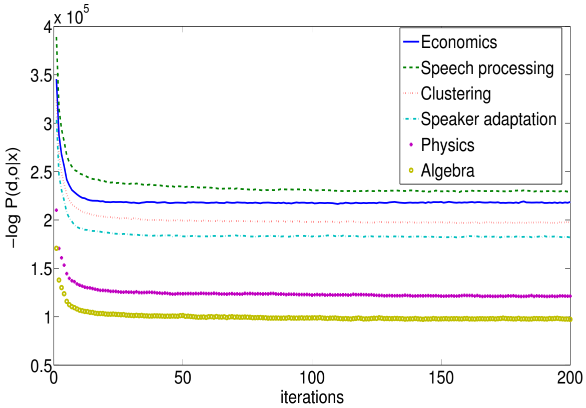

Figure 2: The negative log posterior probability of the latent variables $\boldsymbol{d}$ and $\boldsymbol{o}$ as a function of iteration obtained by the FullDP model for each lecture.

Phoneme segmentation We first evaluate our model on the task of phone segmentation for the six lectures. We use a speech recognizer to produce phone forced alignments for each lecture. The phone segmentation embedded in the forced alignments is then treated as the gold standard to which we compare the segmentation our model generates. We follow the suggestion of (Scharenborg et al., 2010) and use a 20-ms tolerance window to compute the F1 score of all proposed phone segmentations.

Table 2 presents the F1 scores achieved by different systems. Because the -AM and -NC systems do not do inference over acoustic segmentations, we compare the phoneme-segmentation performance of each full system to its performance at initialization. Recall that the Full50 system is initialized using the output of a Hierarchical HMM (HHMM), and the FullDP system is initialized using DPHMM.

From Table 2 we can see that the two -AM systems achieve roughly the same segmentation performance for the first four lectures. Aside from using the boundary elimination method described in (Lee and Glass, 2012), these two systems are trained independently. The table shows that the two initialization systems achieve roughly the same segmentation

Lecture topic	Full50	HHMM	FullDP	DPHMM
Economics	74.4	74.6	74.6	75.0
Signal processing	76.2	76.0	76.0	76.3
Clustering	76.6	76.6	77.0	76.9
Speaker adaptation	76.5	76.9	76.7	76.9
Physics	75.9	74.9	75.7	75.8
Linear algebra	75.5	73.8	75.5	75.7

Table 2: The F1 scores for the phone segmentation task obtained by the full systems and their corresponding initialization systems.

Lecture topic	Full50	-AM	-NC	FullDP	-AM	-NC
Economics	15.4	15.4	14.5	16.1	14.9	13.8
Signal processing	17.5	16.4	12.1	18.3	17.0	14.5
Clustering	16.7	18.1	15.9	18.4	16.9	15.2
Speaker adaptation	17.3	17.4	15.4	18.7	17.6	16.2
Physics	17.7	17.9	15.6	20.0	18.0	15.2
Linear algebra	17.9	17.5	15.4	20.0	17.0	15.6

Table 3: F1 scores for word segmentation obtained by the full systems and their ablated systems.

performance for the first four lectures. Thus, their narrow performance gap indicates that the two initialization systems may have already found a near-optimal segmentation.

Since our model also looks for the best segmentation in the same hypothesis space, by initializing the boundary variables around the optimum, our model should simply maintain the segmentation. In particular, as shown in Table 2, the full systems also achieve about the same performance as the -AM systems for the first four lectures, with the overall largest performance difference being 0.4%.

It is perhaps more interesting when the initialization system gets stuck at a local optimum. By comparing the performance of the two -AM systems for the **last two** lectures, we can see that the initialization of Full50 converges to local optimums for the two lectures. Nonetheless, as shown in Table 2, the Full50 system is able to improve the given initial segmentation and reach a similar performance to that accomplished by the FullDP and the initialization of the FullDP systems.

Word segmentation In addition to phone segmentation, we also evaluate our model on the task of word segmentation. Similar to how we generate the gold standard segmentation for the previous task, we use a speech recognizer to produce word alignments and obtain the word segmentation for each lecture. We then compare the word segmentation that our systems generate to the gold standard and calculate the F1 scores by using a 20-ms tolerance window.

By comparing the full systems to their -NC counterparts, we can see that the noisy-channel model plays an important role for word segmentation, which resonates with the findings of in (Elsner et al., 2013). Without the capability of modeling phonetic variability, it is difficult, or even impossible, for the -NC systems to recognize word tokens of the same type but with different phonetic realizations.

We can also observe the advantage of joint learning both word-level and phone-level representations by comparing the FullDP system to the corresponding -AM model. On average, the FullDP system outperforms its -AM ablated counterpart by 1.6% on the word segmentation task, which indicates that the top-down word-level information can help refine the phone-level knowledge that the model has learned.

While similar improvements are only observed in two lectures for the Full50 and its -AM version, we believe it's because the Full50 system does not have as much flexibility to infer the phonetic embeddings as the FullDP system. This inflexibility may have

Lecture topic	Full50	-AM	-NC	FullDP	-AM	-NC	P&G 2008	Zhang 2013
Economics	12	4	2	12	9	6	11	**14**
Signal processing	16	16	5	**20**	19	14	15	19
Clustering	**18**	17	9	17	**18**	13	16	17
Speaker adaptation	14	14	8	**19**	17	13	13	**19**
Physics	**20**	14	12	**20**	18	16	17	18
Linear algebra	18	16	11	**19**	17	7	17	16

Table 4: The number of the 20 target words discovered by each system described in Sec. 5, and by the baseline (Park and Glass, 2008), and state-of-the-art system (Zhang, 2013). The best performance achieved for each lecture is shown in bold.

prevented the Full50 system to fully exploit the top-down information for learning.

Finally, note that even though the F1 scores for the word segmentation task are low, we find similar performance reported in Jansen et al. (2013). We would like to raise the question of whether the conventional word segmentation task is a proper evaluation method for an unsupervised model such as the one described in this paper. Our thought is two fold. First, *correct segmentation* is vaguely defined. By choosing different tolerance windows, different segmentation performance is obtained. Second, as we show later, many of the units discovered by our model are linguistically meaningful, although they do not always strictly correspond to *words* (i.e., the units may be morphemes or collocations, etc.). Since these are linguistically meaningful units which should be identified by an unsupervised lexical discovery model, it is not clear what advantage would be gained by privileging words in the evaluation. Nevertheless, we present the word segmentation performance achieved by our model in this paper for future references.

Coverage of words with high TFIDF scores To assess the performance of our model, we evaluate the degree to which it was able to correctly recover the vocabulary used in input corpora. To facilitate comparison with the baseline (Park and Glass, 2008) and state-of-the-art (Zhang, 2013) spoken term discovery systems, we restrict attention to the top 20 highest TFIDF scoring words for each lecture. Note that the set of target words of each lecture were originally chosen in (Park and Glass, 2008) and used as the evaluation set in both Park and Glass (2008) and Zhang (2013). To compare our system directly to previous work, we use the same set of target words to test our model.[3]

Table 4 summarizes the coverage achieved by each variant of our model as well as the two spoken term discovery systems. Note that these two systems differ only in the acoustic features used to represent the input: MFCC features versus more robust Gaussian posteriorgrams, respectively. Both the FullDP and Full50 systems consistently outperform the baseline. Both systems also exceed or perform as well as the state-of-the-art system on most lectures. Furthermore, they do so using the less robust MFCC feature-based acoustic representation.

The results in Table 4 also illustrate the synergistic interactions that occur between the acoustic and symbolic model components. As described in Sec. 5, the -AM systems are identical to the full systems, except they do not include the acoustic model, and, therefore, do not re-segment or relabel the speech after initialization. As Table 4 shows, the Full50 and FullDP models both tend to have coverage which is as good as, or better than, the corresponding models without an acoustic component. This indicates that top-down pressure from the symbolic component can refine bottom-layer PLUs, leading, ultimately, to better lexicon discovery.

Finally, the comparison between Full/-AM models and their -NC components suggests the importance of modeling variability in the realization of phonemes. As we will discuss in the next section, the full model tends to merge multiple sequences of

[3] For the procedure used to identify the *word label* for each lexical unit, we refer the reader to Sec. 3.5.3 of (Lee, 2014) for detailed explanation.

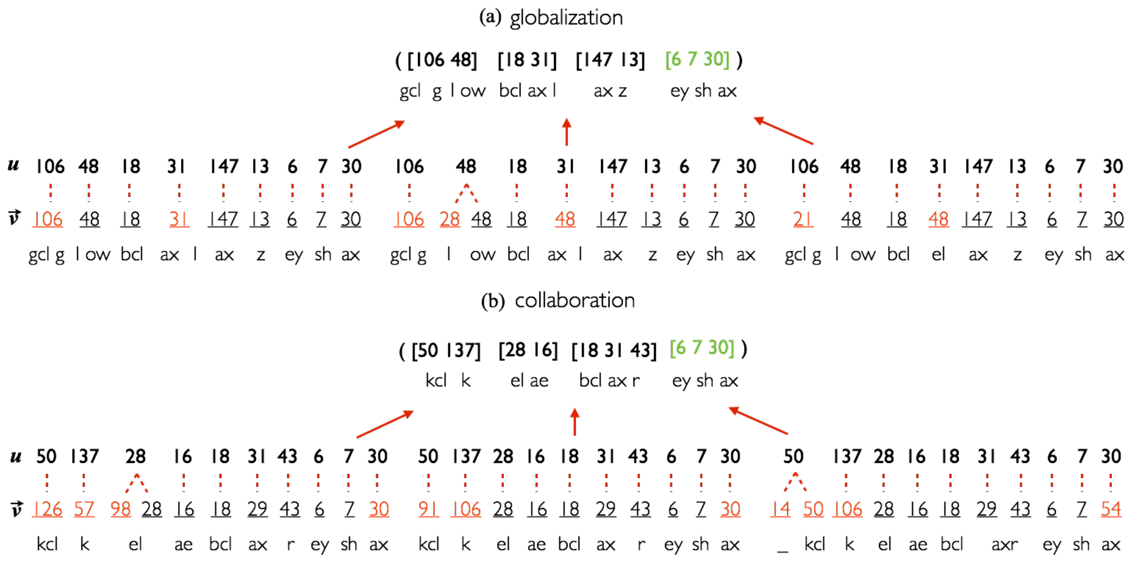
(a) globalization
([106 48] [18 31] [147 13] [6 7 30])
gcl g l ow bcl ax l ax z ey sh ax
u 106 48 18 31 147 13 6 7 30 106 48 18 31 147 13 6 7 30 106 48 18 31 147 13 6 7 30
v 106 48 18 31 147 13 6 7 30 106 28 48 18 48 147 13 6 7 30 21 48 18 48 147 13 6 7 30
gcl g l ow bcl ax l ax z ey sh ax gcl g l ow bcl ax l ax z ey sh ax gcl g l ow bcl el ax z ey sh ax
(b) collaboration
([50 137] [28 16] [18 31 43] [6 7 30])
kcl k el ae bcl ax r ey sh ax
u 50 137 28 16 18 31 43 6 7 30 50 137 28 16 18 31 43 6 7 30 50 137 28 16 18 31 43 6 7 30
v 126 57 98 28 16 18 29 43 6 7 30 91 106 28 16 18 29 43 6 7 30 14 50 106 28 16 18 29 43 6 7 54
kcl k el ae bcl ax r ey sh ax kcl k el ae bcl ax r ey sh ax _ kcl k el ae bcl axr ey sh ax

Figure 3: The bottom-layer PLUs $\vec{v}$ and the top-layer PLUs $\boldsymbol{u}$ as well as the word-internal structure that the FullDP system discovered for three instances of the words (a) *globalization* and (b) *collaboration*. We also include phoneme transcriptions (derived by manual inspection of spectrograms), for clarity.

bottom-layer PLUs into single lexical items which share a single top-layer PLU sequence. The results in Table 4 confirm this: When the model does not have the option of collapsing bottom-layer PLU sequences, word discovery degrades considerably.

Examples and qualitative analyses To provide intuition about the model behavior, we present several examples and qualitative analyses. Figure 3 illustrates the model's representation of two words which appeared frequently in the economics lecture: *globalization* and *collaboration*. The figure shows (i) the bottom-layer PLUs which the model assigned to three instances of each word in the training corpus, (ii) the alignments between these bottom-layer PLUs and the top-layer PLU sequence corresponding the model's lexical representation of each word, (iii) the decomposition of each word-like unit into sub-word-like units, which are denoted as bracketed sequences of PLUs (e.g., [70 110 3]), and (iv) a hand annotated phonemic transcription.

The first thing to note is the importance of the noisy-channel component in normalizing variation across word tokens. In Figure 3-(a) the model has inferred a different sequence of bottom-layer PLUs for each spoken instance of *globalization*. PLUs which vary between the three instances are highlighted in red. The model was able to map these units to the single sequence of top-layer PLUs associated with the lexical item. Similar remarks hold for the word *collaboration* in Figure 3-(b). This suggests that acoustic variability between segments led the model to infer different bottom-layer PLUs between word tokens, but this variability was correctly normalized by the noisy-channel component.

The second thing to note is the large amount of variability in the granularity of stored sub-word-like units (bracketed PLU sequences). The model allows sub-word-like units to consist of any sequence of PLUs, without further constraint. Figure 3 shows that the model makes use of the flexibility, representing linguistic structure at a variety of different scales. For example, the initial sub-word-like unit of *collaboration* groups together two PLUs corresponding to a single phoneme /k/. Other sub-word-like units correspond to syllables. Still others capture morphological structure. For example, the final sub-word-like unit in both words (highlighted in green) corresponds to the combination of suffixes *-ation*—a highly productive unit in English (O'Donnell, 2015).[4]

[4] The reader may notice that the lexical representations are missing a final /n/ phoneme. Manual examination of spectro-

Transcription	Discovered lexical units	|Word|
/iy l iy/ (really, willy, billion)	[35] [31 4]	68
/ey sh ax n/ (innovation, imagination)	[6 7 30] [49]	43
/ax bcl ax l/ (able, cable, incredible)	[34 18] [38 91]	18
discovered	[26] [70 110 3] [9 99] [31]	9
individual	[49 146] [34 99] [154] [54 7] [35 48]	7
powerful	[50 57 145] [145] [81 39 38]	5
open university	[48 91] [4 67] [25 8 99 29] [44 22] [103 4]	4
the arab muslim world	[28 32] [41] [67] [25 35] [1 27] [13 173] [8 139] [38 91]	2

Table 5: A subset of the lexical units that the FullDP system discovers for the economics lecture. The number of independent speech segments that are associated with each lexical unit is denoted as |Word|.

Lexical units also exhibit variability in granularity. Table 5 shows a subset of lexical units discovered by the FullDP system for the economics lecture. Each entry in the table shows (i) the decomposition of each word-like unit into sub-word-like units, (ii) a phonemic transcription of the unit, and (iii) the number of times each lexical unit was used to label a segment of speech in the lecture (denoted as |Word|). The lexical units displayed in the table correspond, for the most part, to linguistic units. While there are a few cases, such as /iy l iy/, where the model stored a sequence of phones which does not map directly onto a linguistic unit such as a syllable, morpheme, or word, most stored units do correspond to intuitively plausible linguistic constituents.

However, like sub-word-like units, there is variability in the scale of the linguistic structure which they capture. On one hand, the model stores a number of highly reusable smaller-than-word units, which typically correspond to morphemes or highly frequent syllables. For example, the sequences /ax bcl ax l/ and /ey sh ax n/ correspond to the productive suffix *-able* and suffix combination *-ation* (O'Donnell, 2015). On the other hand, the model also stores lexical units which correspond to words (e.g., *powerful*) and multi-word collocations (e.g., *the arab muslim world*). Figure 4 shows an analysis of stored lexical units for each lecture, plotting the proportion of stored items which map onto sub-words, words, and multi-word collocations for each.

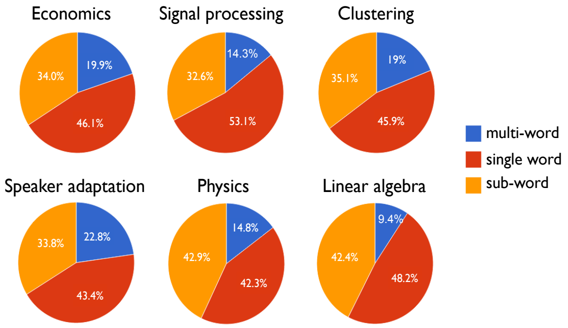

Figure 4: The proportions of the word tokens the FullDP system generates for each lecture that map to *sub-words*, *single words*, and *multi-words*.

Why does the model choose to store some units and not others? Like many approaches to lexicon learning (Goldwater, 2006; De Marcken, 1996c; Johnson et al., 2006), our model can be understood as balancing a tradeoff between productivity (i.e., computation) and reuse (i.e., storage). The model attempts to find a set of lexical units which explain the distribution of forms in the input, subject to two opposing simplicity biases. The first favors smaller numbers of stored units. The second favors derivations of observed utterances which use fewer computational steps (i.e., using small number of lexical items). These are opposing biases. Storing larger lexical units, like *the arab muslim world*, leads to simpler derivations of individual utterances, but a larger lexicon. Storing smaller lexical units, like the suffix *-able*, leads to a more compact lexicon, but more complex derivations individual utterances.

grams revealed two likely reasons for this. First, the final PLU 30 is likely to be a nasalized variant of /ə/, thus encoding some portion of the following /n/ phoneme. Second, across these word instances there is a great deal of acoustic variation between the acoustic realizations of the consonant /n/. It is unclear at present whether this variation is systematic (e.g., co-articulation with the following word), or simply noise.

Smaller units are favored when they are used across a large variety of relatively infrequent contexts. For example, *-ation* appears in a large number of input utterances, but often as part of words which themselves are relatively infrequent (e.g., *conversation*, *reservation*, *innovation*, and *foundation* which appear 2, 3, 4, and 2 times respectively). Larger units will be favored when a combination of smaller units appears more frequently than would be predicted by considering their probabilities in isolation. For example, the model stores the words *globalization* and *collaboration* in their entirety, despite also storing the suffix combination *-ation*. These words occur 25 and 21 times respectively in the lecture, which is a greater number of times than would be expected merely by considering the words sub-parts. Thus, the fact that the model stores a variety of lexical units at different granularities is expected.

7 Conclusion and Future Work

In this paper, we have presented a probabilistic framework for inferring hierarchical linguistic structures from acoustic signals. Our approach is formulated as an integration of adaptor grammars, a noisy-channel model, and an acoustic model. Comparison of the model with lesioned counterparts suggested that our model takes advantage of synergistic interactions between phonetic and lexical representations. The experimental results also indicate that modeling phonetic variability may play a critical role in inferring lexical units from speech.

While the noisy-channel model has demonstrated an ability to normalize phonetic variations, it has its limitations. In the future, we plan to investigate alternatives that more accurately capture phonetic variation. We also plan to explore grammars that encode other types of linguistic structures such as collocation of lexical and morphological units.

Acknowledgements

The authors would like to thank the anonymous reviewers and the action editor of this paper, Eric Fosler-Lussier, for helpful comments. Furthermore, the authors would also like to thank Mark Johnson for making his implementation of adaptor grammars publicly available and for answering detailed questions about the model and his implementation.

References

Guillaume Aimetti. 2009. Modelling early language acquisition skills: Towards a general statistical learning mechanism. In *Proceedings of EACL: Student Research Workshop*, pages 1–9.

Shlomo Argamon, Navot Akiva, Amihood Amir, and Oren Kapah. 2004. Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of the 20th international conference on Computational Linguistics*.

M Bacchiani and M Ostendorf. 1999. Joint lexicon, acoustic unit inventory and model design. *Speech Communication*, 29(24):99 – 114.

Benjamin Börschinger and Mark Johnson. 2014. Exploring the role of stress in Bayesian word segmentation using adaptor grammars. *Transactions of ACL*, 2:93 – 104.

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

Michael R. Brent. 1999b. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, 3(8):294–301, August.

Timothy A. Cartwright and Michael R. Brent. 1994. Segmenting speech without a lexicon: Evidence for a bootstrapping model of lexical acquisition. In *Proceedings of the 16th Annual Meeting of the Cognitive Science Society*.

Siddhartha Chib and Edward Greenberg. 1995. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.

Cheng-Tao Chung, Chun-an Chan, and Lin-shan Lee. 2013. Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization. In *Proceedings of ICASSP*, pages 8081–8085.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1).

Steven B. Davis and Paul Mermelstein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366.

Carl De Marcken. 1996a. Linguistic structure as composition and perturbation. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 335–341. Association for Computational Linguistics.

Carl De Marcken. 1996b. The unsupervised acquisition of a lexicon from continuous speech. Technical Report

AI-memo-1558, CBCL-memo-129, Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Carl De Marcken. 1996c. *Unsupervised Language Acquisition*. Ph.D. thesis, Massachusetts Institute of Technology.

Micha Elsner, Sharon Goldwater, Naomi Feldman, and Frank Wood. 2013. A joint learning model of word segmentation, lexical acquisition, and phonetic variability. In *Proceedings of EMNLP*, pages 42–54.

T.S. Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *Ann. Statist*, 1(2):209–230.

Jenny Rose Finkel, Christopher D. Manning, and Andrew Y. Ng. 2006. Solving the problem of cascading errors: Approximate Bayesian inference for linguistic annotation pipelines. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 618–626.

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.

Michael C. Frank, Joshua B. Tenenbaum, and Edward Gibson. 2013. Learning and long-term retention of large-scale artificial languages. *Public Library of Science (PloS) ONE*, 8(4).

James Glass, Timothy J. Hazen, Lee Hetherington, and Chao Wang. 2004. Analysis and processing of lecture audio data: Preliminary investigations. In *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL*, pages 9–12.

James Glass. 2003. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137 – 152.

John Goldsmith. 2001. Unsupervised learning of the morphology of natural language. *Computational Linguistics*, 27(2):153–198.

John Goldsmith. 2006. An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(4):353–371.

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

Sharon Goldwater. 2006. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

Joshua T. Goodman. 1998. *Parsing Inside-Out*. Ph.D. thesis, Harvard University.

Zellig Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Jahn Heymann, Oliver Walter, Reinhold Haeb-Umbach, and Bhiksha Raj. 2013. Unsupervised word segmentation from noisy input. In *Proceedings of ASRU*, pages 458–463. IEEE.

Aren Jansen, Kenneth Church, and Hynek Hermansky. 2010. Towards spoken term discovery at scale with zero resources. In *Proceedings of INTERSPEECH*, pages 1676–1679.

Aren Jansen, Emmanuel Dupoux, Sharon Goldwater, Mark Johnson, Sanjeev Khudanpur, Kenneth Church, Naomi Feldman, Hynek Hermansky, Florian Metze, Richard C. Rose, et al. 2013. A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition. In *ICASSP*, pages 8111–8115.

Frederick Jelinek. 1976. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64:532 – 556.

Mark Johnson and Katherine Demuth. 2010. Unsupervised phonemic Chinese word segmentation using adaptor grammars. In *Proceedings of COLING*, pages 528–536, August.

Mark Johnson and Sharon Goldwater. 2009a. Improving nonparameteric Bayesian inference: Experiments on unsupervised word segmentation with Adaptor Grammars. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL*, pages 317–325.

Mark Johnson and Sharon Goldwater. 2009b. Improving nonparameteric Bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of NAACL-HLT*, pages 317–325.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. *In Advances in Neural Information Processing Systems*, pages 641–648.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Proceedings of NAACL*, pages 139–146.

Mark Johnson. 2008a. Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, pages 20–27, Columbus, Ohio, June.

Mark Johnson. 2008b. Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In *Proceedings of ACL*, pages 398–406.

Brenden M. Lake, Chia-ying Lee, James R. Glass, and Joshua B. Tenenbaum. 2014. One-shot learning of generative speech concepts. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Soceity*.

Karim Lari and Steve J. Young. 1990. The estimation of stochastic context-free grammars using the inside-outside algorithm. *Computer speech & language*, 4(1):35–56.

Chia-ying Lee and James Glass. 2012. A nonparametric Bayesian approach to acoustic model discovery. In *Proceedings of ACL*, pages 40–49.

Chia-ying Lee, Yu Zhang, and James Glass. 2013. Joint learning of phonetic units and word pronunciations for ASR. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 182–192.

Chia-ying Lee. 2014. *Discovering Linguistic Structures in Speech: Models and Applications*. Ph.D. thesis, Massachusetts Institute of Technology.

Percy Liang, Slav Petrov, Michael I. Jordan, and Dan Klein. 2007. The infinite PCFG using hierarchical Dirichlet processes. In *Processing of EMNLP*, pages 688–697.

Fergus R. McInnes and Sharon Goldwater. 2011. Unsupervised extraction of recurring words from infant-directed speech. In *Proceedings of CogSci*, pages 2006 – 2011.

Daichi Mochihashi, Takeshi Yamada, and Naonori Ueda. 2009. Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In *Proceedings of ACL*, pages 100–108.

Graham Neubig, Masato Mimura, and Tatsuya Kawahara. 2012. Bayesian learning of a language model from continuous speech. *The IEICE Transactions on Information and Systems*, 95(2):614–625.

Timothy J. O'Donnell. 2015. *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. The MIT Press, Cambridge, Massachusetts and London, England.

D. C. Olivier. 1968. *Stochastic Grammars and Language Acquisition Mechanisms*. Ph.D. thesis, Harvard University.

Alex S. Park and James R. Glass. 2008. Unsupervised pattern discovery in speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 16(1):186–197.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Jim Pitman. 1992. The two-parameter generalization of ewens' random partition structure. Technical report, Department of Statistics University of California Berkeley.

Jennifer R. Saffran, Richard N. Aslin, and Elissa L. Newport. 1996a. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, December.

Jennifer R. Saffran, Elissa L. Newport, and Richard N. Aslin. 1996b. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35:606–621.

Odette Scharenborg, Vincent Wan, and Mirjam Ernestus. 2010. Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries. *Journal of the Acoustical Society of America*, 127:1084–1095.

Yaodong Zhang and James Glass. 2009. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams. In *Proceedings of ASRU*, pages 398–403.

Yaodong Zhang, Ruslan Salakhutdinov, Hung-An Chang, and James Glass. 2012. Resource configurable spoken query detection using deep Boltzmann machines. In *Proceedings of ICASSP*, pages 5161–5164.

Yaodong Zhang. 2013. *Unsupervised speech processing with applications to query-by-example spoken term detection*. Ph.D. thesis, Massachusetts Institute of Technology.