18

A factored language model for prosody dependent speech recognition

Ken Chen, Mark A. Hasegawa-Johnson, & Jennifer S. Cole
University of Illinois at Urbana-Champaign
U.S.A.

1. Introduction

Prosody refers to the suprasegmental features of natural speech (such as rhythm and intonation) that are used to convey linguistic and paralinguistic information (such as emphasis, intention, attitude, and emotion). Humans listening to natural prosody, as opposed to monotone or foreign prosody, are able to understand the content with lower cognitive load and higher accuracy (Hahn, 1999). In automatic speech understanding systems, prosody has been previously used to disambiguate syntactically distinct sentences with identical phoneme strings (Price et al., 1991), infer punctuation of a recognized text (Kim & Woodland, 2001), segment speech into sentences and topics (Shriberg et al., 2000), recognize the dialog act labels (Taylor et al., 1997), and detect speech disfluencies (Nakatani and Hirschberg, 1994). None of these applications use prosody for the purpose of improving word recognition (i.e., the word recognition module in these applications does not utilize any prosody information). Chen et al. (Chen et al., 2003) proposed a prosody dependent speech recognizer that uses prosody for the purpose of improving word recognition accuracy. In their approach, the task of speech recognition is to find the sequence of word labels $W=(w_1,\ldots,w_M)$ that maximizes the recognition probability:

$$\begin{aligned}[\hat{W}] &= \arg\max\, p(O\,|\,W,P)\,p(W,P) \\ &= \arg\max\, p(O\,|\,Q,H)\,p(Q,H\,|\,W,P)\,p(W,P),\end{aligned} \tag{1}$$

where $P=(p_1,\ldots,p_M)$ is a sequence of prosody labels, one associated with each word, $O=(o_1,\ldots,o_T)$ is a sequence of observed acoustic feature vectors, $Q=(q_1,\ldots,q_L)$ is a sequence of sub-word units, typically allophones dependent on phonetic context, and $H=(h_1,\ldots,h_L)$ is a sequence of discrete ``hidden mode'' vectors describing the prosodic states of each allophone. The combination $[w_m,p_m]$ is called a prosody-dependent word label, the combination $[q_l,h_l]$ is called a prosody-dependent allophone label, $p(O\,|\,Q,H)$ is a prosody-dependent acoustic model, $p(Q,H\,|\,W,P)$ is a prosody-dependent pronunciation model, and $p(W,P)$ is a prosody-dependent language model. In this framework, word and prosody are conditioned on each other and are recognized at the same time. The system described in

equation (1) has the advantage that both the acoustic model and the language model can be potentially improved through their dependence on prosody.

In (Chen et al. 2006), the prosody variable p_m takes 8 possible values composed by 2 discrete prosodic variables: a variable a that marks a word as either ``a'' (pitch-accented) or ``u'' (pitch-unaccented), and a variable b that marks a word as ``i,m,f,o'' (phrase-initial, phrase-medial, phrase-final, one-word phrase) according to its position in an intonational phrase. Thus, in this scheme, a prosody-dependent word transcription may contain prosody-dependent word tokens of the form $w_{ab}$. For example, the sentence ``well, what's next,'' uttered as two intonational phrases with two accented words, might be transcribed as ``well$_{ao}$ what's$_{ui}$ next$_{af}$.''

A prosody dependent language model $p(W,P)$ that models the joint probability distribution of concurrent word and prosody sequences, is different from a standard prosody independent language model $p(W)$ in the sense that not only word context but also prosody context affect the prediction of the next possible word and its prosody. This model is useful in at least two respects. First, it can be used to effectively reduce the search space of possible word hypotheses. (Kompe, 1997) have shown that a prosody dependent language model can be used to speed up the word recognition process without sacrificing accuracy. Second, it is potentially useful in improving word recognition accuracy. Waibel (Waibel, 1988) reports that prosodic knowledge sources, when added to a phonetic speaker-independent word hypothesizer, are able to reduce the average rank of the correct word hypothesis by a factor of 3. Arnfield (Arnfield, 1994) gives an example in his dissertation: the words ``witch'' and ``which'', having identical acoustic observations, can be distinguished prosodically (``witch'' is more likely to be accented than is ``which'' because it is a content word while ``which'' is a function word). The word to be predicted is more likely to be ``witch'' instead of ``which'' if an accent is predicted from the current word-prosody context. In the results reported by (Chen et al., 2006), a prosody dependent language model can significantly improve word recognition accuracy over a prosody independent language model, given the same acoustic model.

N-gram models can be conveniently used for prosody dependent language modeling. The n-gram probabilities are estimated from their maximum likelihood estimators (the relative frequency count of the n-grams). For example, the bigram probability $p(w_j, p_j \mid w_i, p_i)$ (the probability of observing token $[w_j, p_j]$ given token $[w_i, p_i]$) can be estimated using the following equation:

$$p(w_j, p_j \mid w_i, p_i) = \frac{n(w_j, p_j, w_i, p_i)}{n(w_i, p_i)}, \tag{2}$$

where $n(\cdot)$ is the number of the n-grams observed in the training set. Equation (2) treats each prosody dependent word token $[w_j, p_j]$ as a distinct unit, resulting in a recognizer that has $|p|$ times larger vocabulary size than does a standard prosody independent recognizer (where $|p|$ is the number of options for tag p_i). If any word-prosody combination can occur in English, the number of prosody dependent n-grams is equal to $|p|^n$ times the number of prosody independent n-grams. In practice, the number of possible prosody

dependent n-grams increases by far less than $|p|^n$ times, because a considerable amount of prosody dependent n-grams never occur in natural English. Nevertheless, the number of possible prosody dependent n-grams still greatly increases as $|p|$ increases due to the prosody variation induced by high level contextual information and by different speaking styles. Hence, robust estimation of prosody dependent language modeling using equation (2) requires an increasingly large amount of prosodically labeled data which are normally expensive to acquire. When the size of training text is limited, increasing $|p|$ decreases the trainability of the n-gram models and reduces the consistency between the training and test text: the accuracy of the estimated probability mass functions (PMFs) decreases due to the prosody induced data sparseness and the number of possible unseen prosody dependent n-grams increases.

In this chapter, we propose to improve the robustness of prosody dependent language modeling by utilizing the dependence between prosody and syntax. There is evidence indicating that syntax is a strong conditioning factor for prosody. For example, conjunctions (e.g., ``but",``so") occur more frequently at phrase initial positions than at phrase medial or final positions in fluent speech; content words (e.g., nouns) have much higher probability of being accented than function words (e.g., prepositions, articles). In a corpus based study, Arnfield (Arnfield, 1994) proved empirically that although differing prosodies are possible for a fixed syntax, the syntax of an utterance can be used to generate an underlying ``baseline" prosody regardless of actual words, semantics or context. The bigram models developed by Arnfield were able to predict prosody from parts-of-speech with a high accuracy (91% for stress presence prediction). The experiments conducted by (Hirschberg, 1993) and (Chen et al., 2004) also indicate that parts-of-speech can predict the presence of pitch accent with accuracies of around 82%-85% on the Radio New Corpus.

This chapter is organized as following: Section 2 reviews previous research on factored language models and provides a Bayesian network view of spoken language that further explains our motivation, Section 3 describes our methods for creating prosody dependent factored language models, Section 4 reports our experiments on the Radio News Corpus and discusses results, and conclusions are given in Section 5.

2.Background: Factored Language Models

2.1 Previous work

The objective of a statistical language model is to accurately predict the next word w_j from current history $h_j = [w_0, \ldots w_{j-1}]$. In the past two decades, enormous efforts have been reported in the literature to find the factors in h_j that best predict w_j (Rosenfeld, 2000) including the use of syntactic and semantic information extracted from h_j (Khudanpur & Wu, 2000; Bellegarda, 2000). Language modeling for speech recognition has been shown to be a difficult task due to the many sources of variability existing in spoken language including disfluency, sentence fragments, dialect, and stylistic and colloquial language use. The existence of these intrinsic properties of spoken language (which are quit different from written language) have forced researchers to expand the space h_j to include additional streams of knowledge.

One example is the system proposed by Heeman and Allen (Heeman and Allen, 1999), in which the word sequences and parts-of-speech (POS) sequences are modeled jointly and recognized simultaneously in a unified framework:

$$\begin{aligned}[\tilde{W},\tilde{S}] &= \arg\max\, p(O \mid W,S)\,p(W,S) \\ &\approx \arg\max\, p(O \mid W)\,p(W,S).\end{aligned} \tag{3}$$

The language model $p(w_j,s_j \mid W_{0,j-1},S_{0,j-1})$ can be factored into two component language models, which makes it possible to utilize the syntactic knowledge encoded in the joint history of word and POS to improve the predictability of the next word (and its POS):

$$\begin{aligned}&p(w_j,s_j \mid W_{0,j-1},S_{0,j-1}) \\ &= p(w_j \mid W_{0,j-1},S_{0,j-1},s_j)\,p(s_j \mid W_{0,j-1},S_{0,j-1}),\end{aligned} \tag{4}$$

where s_j is the POS of w_j, $W_{0,j-1}$ is the word history up to w_{j-1}, and $S_{0,j-1}$ is the POS history up to s_{j-1}. Heeman used decision trees to cluster the word and POS history into equivalence classes. This multi-stream POS-based language model $p(W,S)$ achieved a 7% reduction of word perplexity over the single stream word-based n-gram language model $p(W)$ and improved the prediction of word and POS simultaneously. Heeman further extended this multi-stream language modeling frame work to include more knowledge sources (e.g., intonational phrase boundaries, speech repairs, discourse markers) and found that the inter-dependence among these knowledge sources can further improve the quality of the language model for the modeling of conversational spoken language.

In a different context, Kirchhoff (Kirchhoff et al., 2003) applied the idea of multi-stream language modeling to handle the morphological complexity in Arabic text, where she modeled multiple streams of word features such as the morphological class (m_i), patterns (p_i) and roots (r_i) in place of the single stream of words. For example, represent $w_i = (r_i, p_i, m_i)$,

$$\begin{aligned}&p(w_i \mid w_{i-1},w_{i-2}) \\ &= p(r_i, p_i, m_i \mid r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &= p(r_i \mid p_i, m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\quad p(p_i \mid m_i, r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}) \\ &\quad p(m_i \mid r_{i-1}, p_{i-1}, m_{i-1}, r_{i-2}, p_{i-2}, m_{i-2}).\end{aligned} \tag{5}$$

The three factored probability functions in equation (5) can be modeled individually using n-grams or other modeling techniques. Since each word feature has much smaller cardinality than the word vocabulary, and is less fractured by the nuances of morphological variation, this factored language model can effectively help reduce the data sparseness in dialectal Arabic.

The language models we are proposing can be viewed as an extension to these previous works. Rather than modeling POS explicitly in the language model, we propose to model prosody explicitly while using POS implicitly to reduce the data sparseness induced by prosody. We argue that this method of modeling prosody makes the acoustic models and language models fuse more tightly through their interaction with prosody and brings the potential of improving both word recognition and prosody recognition performance.

2.2 A Bayesian Network View for Spoken Language

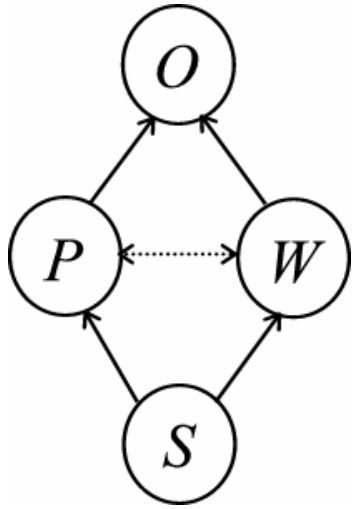

Figure 1. A Bayesian network representing the complex relationship among the acoustic observation sequence (O), word sequence (W), prosody sequence (P) and syntax sequence (S) of an utterance.

To better understand our reasoning behind the idea of prosody dependent speech recognition, we plot in Fig. 1 the complex relationship among the sequences of acoustic observations O, words W, prosody P and syntax S for an arbitrary utterance in terms of a Bayesian Network. The dependence of O over P is well defined because it is well known that prosody affects the acoustic realization of words in systematic ways. For example, unaccented vowels tend to be centralized and reduced in a function word, accented vowels tend to be longer and less subject to coarticulatory variation (Cho, 2001); accented consonants are produced with greater closure duration (DeJong, 1995), greater linguopalatal contact (Fougeron & Keating, 1997), longer voice onset time, and greater burst amplitude (Cole et al., 2007). Conditioning O over both P and W brings us a framework in which prosody induced acoustic variations can be accurately modeled. The dependence of W over S is well-established and has been used to build various types of POS-based language models. The dependence of P over S is supported by the experiments of Arnfield and others, described at the end of Section 1. The inter-dependence between P and W has been depicted by a dashed arrow to express the fact that P can be assumed to be independent of W given S with no knowledge about the pragmatic context (i.e., there is no reason to believe that one noun is more likely to be accented than any other given no pragmatic context). This assumption is useful in our later derivation in Section 3.

Modeling W and P jointly in this prosody dependent framework creates a new search space in which the candidate word sequences are weighted in terms of their conformability with natural prosody. An information-theoretic analysis in (Chen & Hasegawa-Johnson, 2004; Hasegawa-Johnson et al, 2005; Chen et al., 2006) showed that it is possible for a

prosody-dependent speech recognizer to improve word recognition accuracy even if the acoustic model and the language model do not separately lead to improvements. Even if prosody does not improve the recognition of words in isolation, the likelihood of the correct sentence-level transcription may be improved by a language model that correctly predicts prosody from the word string, and an acoustic model that correctly predicts the acoustic observations from the prosody. In their experiments on the Radio News Corpus (Chen et al., 2006), as large as 11% word recognition accuracy improvement over a prosody independent speech recognizer was achieved by a prosody dependent recognizer that has comparable total parameter count.

3. Method

In this section, we propose an approach that creates prosody dependent factored language models by utilizing the dependence between prosody and syntax. For notational convenience and clarity, we used bigram models for our derivation. The equations presented in this section can be easily extended to higher order n-gram models.

3.1 Prosody Dependent Factored Language Model

The semi prosody dependent bigram probability (the probability of observing a word w_j given the previous prosody dependent word label $[w_i, p_i]$) can be calculated from the prosody independent bigram probability $p(w_j \mid w_i)$ using the following equation:

$$\begin{aligned} & p(w_j \mid w_i, p_i) \\ &= \frac{p(p_i, w_j \mid w_i)}{p(p_i \mid w_i)} \\ &= \frac{p(p_i \mid w_j, w_i)p(w_j \mid w_i)}{p(p_i \mid w_i)} \\ &\approx \frac{\sum_{s_i,s_j} p(p_i \mid s_i, s_j)p(s_i, s_j \mid w_i, w_j)p(w_j \mid w_i)}{\sum_{w_j}\sum_{s_i,s_j} p(p_i \mid s_i, s_j)p(s_i, s_j \mid w_i, w_j)p(w_j \mid w_i)}, \end{aligned} \tag{6}$$

where s_i and s_j are the POS of w_i and w_j respectively. The approximation in equation (6) assumes that p_i (the prosody on the previous word) is dependent on the POS context but independent of the actual word context:

$$p(p_i \mid s_i, s_j) \approx p(p_i \mid s_i, s_j, w_i, w_j). \tag{7}$$

Similarly, the prosody dependent bigram probability (the probability of observing a prosody dependent word token $[w_j, p_j]$ given the previous prosody dependent word token $[w_i, p_i]$) can be calculated from the semi prosody dependent bigram probability $p(w_j \mid w_i, p_i)$:

$$
\begin{aligned}
& p(w_j, p_j \mid w_i, p_i) \\
&= p(p_j \mid w_j, w_i, p_i) p(w \mid w_i, p_i) \\
&= \sum_{s_i, s_j} p(p_j \mid s_j, s_i, w_j, w_i, p_i) p(s_j, s_i \mid w_j, w_i, p_i) p(w_j \mid w_i, p_i) \\
&\approx \sum_{s_i, s_j} p(p_j \mid s_j, s_i, p_i) p(s_j, s_i \mid w_j, w_i) p(w_j \mid w_i, p_i).
\end{aligned} \tag{8}
$$

The following approximations are required in deriving equation (8):

$$p(p_j \mid s_i, s_j, p_i) \approx p(p_j \mid s_i, s_j, w_i, w_j, p_i), \tag{9}$$

and

$$p(s_i, s_j \mid w_i, w_j) \approx p(s_i, s_j \mid w_i, w_j, p_i). \tag{10}$$

Equation (9) assumes that prosody is dependent on its syntactic context represented by the POS of current word and the previous word but independent of the actual words. Equation (10) assumes that prosody does not affect the probability distribution of POS given the actual word context. This assumption is plausible except for the cases where prosody is used to resolve syntactic ambiguities (Price et al., 1991). In this chapter, we assume that the use of prosody to resolve POS ambiguity is statistically rare.

Equations (6) and (8) provide an approach to calculate the prosody dependent bigram probability $p(w_j, p_j \mid w_i, p_i)$ based on the regular prosody independent bigram probability $p(w_j \mid w_i)$ and three additional probability mass functions: $p(p_i \mid s_i, s_j)$, $p(p_j \mid s_i, s_j, p_i)$, and $p(s_i, s_j \mid w_i, w_j)$. $p(s_i, s_j \mid w_i, w_j)$ describes the stochastic mapping between a word pair and the associated POS pair. In most cases, this probability is a delta function, meaning that a word pair can only be associated with a unique POS pair. In a few cases, it is possible for a word pair to have more than one associated POS pairs. The probability mass functions $p(p_i \mid s_i, s_j)$ and $p(p_j \mid s_i, s_j, p_i)$ describe the inter-dependence between prosody and parts-of-speech, and can be very robustly estimated from a small database due to the small cardinality of the POS set and the prosody set. Note that equation (8) is possibly more accurate than equation (6) because the approximations are made only in the numerator while equation (6) has approximations in both numerator and denominator.

3.2 Methods for Smoothing the Language Models

Two popular techniques can be used to smooth the resulting language model: the backoff scheme and linear interpolation. When a prosody dependent bigram can not be estimated from the training data, it can be backed off to a prosody dependent unigram using Katz's backoff scheme (Katz, 1987):

$$p_b(w_j, p_j \mid w_i, p_i) = \begin{cases} d_r p(w_j, p_j \mid w_i, p_i), & \text{if exists} \\ b(w_i, p_i) p(w_j, p_j), & \text{else} \end{cases}, \tag{11}$$

where $0 < d_r \leq 1$ is a constant discount ratio and the backoff weight $b(w_i, p_i)$ is computed to ensure that the bigram probabilities conditioned on $[w_i, p_i]$ sum up to 1:

$$b(w_i, p_i) = \frac{1 - \sum_{j \in B} p(w_j, p_j \mid w_i, p_i)}{1 - \sum_{j \in B} p(p_j \mid w_j)}, \tag{12}$$

where B is the set of all prosody dependent word labels $[w_j, p_j]$ whose bigram probabilities can be calculated from equations (6) and (8).

The bigram probabilities calculated from equations (6) and (8) can be interpolated with the bigram probabilities estimated directly from the data (equation (2)). Let p_c be the probabilities calculated by equation (6) and (8), and p_m the probabilities estimated by equation (2), the interpolated probability p_i can be obtained using:

$$\begin{aligned} & p_i(w_j, p_j \mid w_i, p_i) \\ & = \lambda p_c(w_j, p_j \mid w_i, p_i) + (1 - \lambda) p_m(w_j, p_j \mid w_i, p_i), \end{aligned} \tag{13}$$

where λ is a constant weight optimized using an EM algorithm to minimize the cross entropy of the interpolated language model over an independent development-test set.

3.3 Joint Perplexity and Word Perplexity

The quality of a standard prosody independent language model $p(W)$ can be measured by its perplexity E over a test set $T = [w_0, w_1, \ldots, w_N]$:

$$E(T) = 2^{H(T)}, \tag{14}$$

where the cross-entropy $H(T)$ can be calculated as:

$$H(T) = -\frac{1}{N} \sum_{k=1}^{N} \log_2 p(w_k \mid w_{k-1}). \tag{15}$$

Similarly, the quality of a prosody dependent language model $p(W,P)$ can be measured by its perplexity E_p over the test set $T_p = [w_0, p_0, w_1, p_1, \ldots, w_N, p_N]$ that contains the same word sequence as T does but is transcribed prosodically:

$$E_p(T_p) = 2^{H_p(T_p)}, \tag{16}$$

where $H_p(T_p)$ can be calculated as:

$$H_p(T_p) = -\frac{1}{N} \sum_{k=1}^{N} \log_2 p(w_k, p_k \mid w_{k-1}, p_{k-1}). \tag{17}$$

To avoid confusion, we name E the Word Perplexity, and E_p the Joint Perplexity. Obviously, E and E_p are not directly comparable because they are calculated over different hypothesis spaces: E is an estimate of how many possible words can appear in the next spot given current word history, while E_p is an estimate of how many possible prosody dependent word tokens can appear in the next spot given current word and prosody history.

To directly compare the quality of a prosody dependent language model with that of a prosody independent language model, we need to compute the word perplexity for the prosody dependent language model. Note that equation (15) can be expanded as

$$H(T) = -\frac{1}{N}\sum_{k=1}^{N}\log_2 \frac{\sum_{w_{0,k-1}}\sum_{P_{0,k}} p(W_{0,k}P_{0,k})}{\sum_{w_{0,k-2}}\sum_{P_{0,k-1}} p(W_{0,k-1}P_{0,k-1})}, \quad (18)$$

where $W_{0,k} = [w_0, w_1, \ldots w_k]$, $P_{0,k}$ includes all possible prosody paths that can be assigned to $W_{0,k}$, and $p(W_{0,k}P_{0,k})$ is calculated using the estimated prosody dependent language model. Note that equation (18) can be computed efficiently using the forward algorithm, one of the standard algorithms for HMM.

4. Experiments and Results

4.1 The Corpus

To train prosody dependent speech recognizers, a large prosodically labeled speech database is required. The Boston University Radio News Corpus is one of the largest corpora designed for study of prosody (Ostendorf et al., 1995). The corpus consists of recordings of broadcast radio news stories including original radio broadcasts and laboratory broadcast simulations recorded from seven FM radio announcers (4 male, 3 female). Radio announcers usually use more clear and consistent prosodic patterns than non-professional readers, thus the Radio News Corpus comprises speech with a *natural but controlled* style, combining the advantages of both read speech and spontaneous speech. In this corpus, a majority of paragraphs are annotated with the orthographic transcription, phone alignments, part-of-speech tags and prosodic labels. The part-of-speech tags used in this corpus are the same as those used in the Penn Treebank. This tag set includes 47 parts-of-speech: 22 open class categories, 14 closed class categories and 11 punctuation labels. Part-of-speech labeling is carried out automatically using the BBN tagger. The tagger uses a bigram model of the tag sequence and a probability of tag given word taken from either a dictionary or, in the case of an unknown word, based on features of the word related to endings, capitalization and hyphenation. The tagger was trained on a set of Wall Street Journal sentences that formed part of the Penn Treebank corpus. For the labnews stories (a subset of the Radio New Corpus recorded without noise in a phonetics laboratory), only 2% of the words were incorrectly labeled.

The prosodic labeling system represents prosodic phrasing, phrasal prominence and boundary tones, using the Tones and Break Indices (ToBI) system for American English (Beckman & Ayers, 1994). The ToBI system labels pitch accent tones, phrase boundary tones, and prosodic phrase break indices. Break indices indicate the degree of decoupling between each pair of words; intonational phrase boundaries are marked by a break index of 4 or higher. Tone labels indicate phrase boundary tones and pitch accents. Tone labels are constructed from the three basic elements H, L, and !H. H and L represent high tone, low tone respectively, while !H represents a high tone produced at a pitch level that is stepped down from the level of the preceding high tone. There are four primary types of intonational phrase boundary tones: L-L%, representing the pitch fall at the end of a declarative phrase or sentence; H-L%, representing a fall or plateau at a mid-level pitch such as occurs in the middle of a longer declarative dialog turn; H-H%, representing the canonical, upward pitch contour at the end of a yes-no question; and L-H%, representing the low-rising contour found at the end of each non-final item on a list. The contours !H-L% and !H-H% are down-stepped variants that may occur following a H* pitch accent and are less frequently

observed. Seven types of accent tones are labeled: H*, !H*, L+H*, L+!H*, L*, L*+H and H+!H*. The ToBI system has the advantage that it can be used consistently by labelers for a variety of styles. For example, if one allows a level of uncertainty in order to account for differences in labeling style, it can be shown that the different transcribers of the Radio News Corpus agree on break index with 95% inter-transcriber agreement (Ostendorf et al., 1995). Presence versus absence of pitch accent is transcribed with 91% inter-transcriber agreement.

In the experiments we report in this chapter, the original ToBI labels are simplified: accents are only distinguished by presence versus absence, word boundaries are distinguished by those in intonational phrase-final position, and those that are medial in an intonational phrase. Applying this simplification, we create prosody dependent word transcriptions in which a word can only have 4 possible prosodic variations: unaccented phrase medial (``um''), accented phrase medial (``am''), unaccented phrase final (``uf'') and accented phrase final (``af'').

4.2 Perplexity

The prosodically labeled data used in our experiments consist of 300 utterances, 24944 words (about 3 hours of speech sampled at 16Khz) read by five professional announcers (3 female, 2 male) containing a vocabulary of 3777 words. Training and test sets are formed by randomly selecting 85% of the utterances for training, 5% of the utterances for development test and the remaining 10% for testing (2503 words).

We first measured the quality of the language models in terms of their perplexity on the test set. Four language models are trained from the same training set: a standard prosody independent backoff bigram language model LPI, a prosody dependent backoff bigram language model LPDM computed using equation (2), a prosody dependent backoff bigram language model LPDC1 computed using equation (8) only, and a model LPDC2 computed using both equation (6) and equation (8). The difference between LPDC1 and LPDC2 is that in LPDC1, the semi prosody dependent bigram probabilities $p(w_j \mid w_i, p_i)$ required by equation (8) are estimated directly from training data using their maximum likelihood estimators; whereas in LPDC2 they are computed from the prosody independent bigram probabilities $p(w_j \mid w_i)$ using equation (6). The models LPDC1 and LPDC2 were linearly interpolated with LPDM using equation (13), with interpolation weights λ optimized over the development-test set. Table I lists the results of this experiment.

	LPI	**LPDM**	**LPDC1**	**LPDC2**
Joint Perp.		340	282	235
Word Perp.	130	60	54	47
Unseen bigrams	931	1244	1103	956
Total bigrams	12100	14461	37373	81950

Table 1. The joint perplexity, word perplexity, number of unseen bigrams in the test set and total number of estimated bigrams in the prosody independent language model (LPI), the prosody dependent language model estimated using the standard ML approach (LPDM) and the prosody dependent language model calculated using the proposed algorithm (LPDC1 and LPDC2).

Compare the performance among the prosody dependent language models: LPDM, LPDC1, and LPDC2. Both LPDC1 and LPDC2 have much smaller joint perplexity than LPDM: the joint perplexity of LPDC1 is 17% less than that of LPDM, while that of LPDC2 is 31% smaller. Factorial modeling increases the number of bigrams whose probabilities can be estimated more accurately than their backed-off unigrams: the number of total estimated bigrams increased by 2 and 7 times respectively in LPDC1 and LPDC2, and the number of unseen bigrams in the test data reduced by around 25%, approaching the number of unseen bigrams in LPI.

To compare the perplexity of the prosody dependent language models with the prosody independent language model LPI, we computed the word perplexity for the prosody dependent language models using equation (18). As can be seen in the third row of Table 1, word perplexity of LPDC2 is reduced by 64% relative to LPI. Note that the word perplexities of the prosody dependent language models are only weakly comparable with that of the prosody independent language models in terms of predicting the word recognition performance. The word recognition power of a prosody dependent language model is prominent only when it is coupled with an effective prosody dependent acoustic model.

4.3 Word Recognition

Encouraged by the great reduction in perplexity, we conducted word and prosody recognition experiments on the same training and test sets. Two acoustic models are used in this experiment: a prosody independent acoustic model API and a prosody dependent acoustic model APD. All phonemes in API and APD are modeled by HMMs consisting of 3 states with no skips. Within each state, a 3 mixture Gaussian model is used to model the probability density of a 32-dimensional acoustic-phonetic feature stream consisting of 15 MFCCs, energy and their deltas. The allophone models in APD contain an additional one-dimensional Gaussian acoustic-prosodic observation PDF which is used to model the probability density of a nonlinearly-transformed pitch stream, as described in (Chen et al, 2004; Chen et al, 2006). API contains monophone models adopted from the standard SPHINX set (Lee, 1990) and is unable to detect any prosody related acoustic effects. APD contains a set of prosody dependent allophones constructed from API by splitting the monophones into allophones according to a four-way prosodic distinction (unaccented medial, accented medial, unaccented final, accented final): each monophone in API has 4 prosody dependent allophonic variants in APD. Allophone models in APD that are split from the same monophone share a single tied acoustic-phonetic observation PDF, but each allophone distinctly models the state transition probabilities and the acoustic-prosodic observation PDF. The APD allophones are therefore able to detect two of the most salient prosody induced acoustic effects: preboundary lengthening, and the pitch excursion over the accented phonemes. The parameter count of the acoustic-phonetic observation PDF (195 parameters per state) is much larger than the parameter count of the acoustic-prosodic observation PDF (2 parameters per state) or the transition probabilities (1 parameter per state); since the acoustic-phonetic parameters are shared by all allophones of a given monophone, the total parameter count of the APD model set is only about 6% larger than the parameter count of API.

Five recognizers are tested: a standard prosody independent recognizer RII using API and LPI, a semi prosody independent recognizer RID using APD and LPI, a prosody dependent

recognizer RDM using APD and LPDM, a prosody dependent recognizer RDC1 using APD and LPDC1, and a prosody dependent recognizer RDC2 using APD and LPDC2. The word recognition accuracy, accent recognition accuracy and intonational phrase boundary recognition accuracy of these recognizers over the same training and test set are reported in Table 2.

	RII	**RID**	**RDM**	**RDC1**	**RDC2**
AM	API	APD	APD	APD	APD
LM	LPI	LPI	LPDM	LPDC1	LPDC2
Word	75.85	76.02	77.29	78.27	77.08
Accent	56.07	56.07	79.59	79.71	80.26
IPB	84.97	84.97	85.06	85.80	86.62

Table 2. Percent word, accent, and intonational phrase boundary recognition accuracy for recognizers RII, RID, RDM, RDC, and RDC2.

Overall, the prosody dependent speech recognizers significantly improve the word recognition accuracy (WRA) over the prosody independent speech recognizer. RDM improved the word recognition accuracy by 1.4% over RII and 1.2% over RID. RDC1 further improved the WRA by 1% over RDM, apparently benefiting from the improved prosody language model LPDC1. The pitch accent recognition accuracy (ARA) and the intonational phrase boundary recognition accuracy (BRA) are also significantly improved. Since RII and RID classify every word as unaccented and every word boundary as phrase-medial, the ARA and BRA listed in RII and RID are the chance levels. RDM showed a great improvement in ARA but only slight improvement in BRA mostly due to the already high chance level 84.97%. RDC2 used the language model LPDC2 that has the smallest perplexity. However, it only achieved improvement over RDM on ARA and BRA (0.7% and 1.5% respectively), but not on WRA. The failure of LPDC2 to outperform the WRA of LPDC1 may not be meaningful: it is well known that perplexity does not always correlate with recognition performance. However, it is possible to speculatively assign some meaning to this result. The flexible class-dependent structure of LPDC2 is able to model a number of prosody-dependent bigrams that is seven times larger than the number observed in the training data (Table I). It is possible that the approximations in equation (6) do not accurately represent the probabilities of all of these bigrams, and that therefore the increased flexibility harms word recognition accuracy.

5. Conclusion

In this chapter, we proposed a novel approach that improves the robustness of prosody dependent language modeling by leveraging the dependence between prosody and syntax. In our experiments on Radio News Corpus, a factorial prosody dependent language model estimated using our proposed approach has achieved as much as 31% reduction of the joint perplexity over a prosody dependent language model estimated using the standard Maximum Likelihood approach. In recognition experiments, our approach results in a 1% improvement in word recognition accuracy, 0.7% improvement in accent recognition accuracy and 1.5% improvement in intonational phrase boundary (IPB) recognition accuracy

over the baseline prosody dependent recognizer. The study in the chapter shows that prosody-syntax dependence can be used to reduce the uncertainty in modeling concurrent word-prosody sequences.

6. Acknowledgment

This work was supported in part by NSF award number 0132900, and in part by a grant from the University of Illinois Critical Research Initiative. Statements in this chapter reflect the opinions and conclusions of the authors, and are not endorsed by the NSF or the University of Illinois.

7. References

Arnfield, S. (1994). Prosody and syntax in corpus based analysis of spoken English, Ph.D. thesis, University of Leeds

Beckman, M. E. and Ayers, G. M. (1994). Guidelines for ToBI Labelling: the Very Experimental HTML Version, http://www.ling.ohio-state.edu/research/phonetics/E ToBI/singer tobi.html

Bellegarda, J. (2000). Exploiting latent semantic information in statistical language modeling, *Proceedings of the IEEE* 88, 8, 1279-1296

Chen, K., Borys, S., Hasegawa-Johnson, M., and Cole, J. (2003). Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries, *Proc. EUROSPEECH*, Geneva, Switzerland

Chen, K. and Hasegawa-Johnson, M. (2004). How prosody improves word recognition, *Proc. ISCA International Conference on Speech Prosody*, Nara, Japan

Chen, K., Hasegawa-Johnson, M., and Cohen, A. 2004. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model, *Proc. ICASSP*, Montreal, Canada

Chen, K., Hasegawa-Johnson, M., Cohen, A., Borys, S., Kim, S., Cole, J., and Choi, J. (2006). Prosody dependent speech recognition on radio news, *IEEE Trans. Speech and Audio Processing* 14(1): 232-245

Cho, T. 2001. Effects of prosody on articulation in English, Ph.D. thesis, UCLA

Cole, J., Kim H., Choi H., & Hasegawa-Johnson M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics* 35: 180-209

DeJong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation, *J. Acoust. Soc. Am* 89, 1, 369-382

Fougeron, C. and Keating, P. (1997). Articulatory strengthening at edges of prosodic domains, *J. Acoust. Soc. Am* 101, 6, 3728-3740

Hahn, L. (1999). Native speakers' reactions to non-native stress in English discourse, Ph.D. thesis, UIUC

Hasegawa-Johnson M., Chen K., Cole J., Borys S., Kim S., Cohen A., Zhang T., Choi J., Kim H., Yoon T., and Chavarria S. (2005). Simultaneous Recognition of Words and Prosody in the Boston University Radio Speech Corpus, *Speech Communication*, 46(3-4): 418-439

Heeman, P. and Allen, J. (1999). Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog, *Computational Linguistics* 25, 4

Hirschberg, J. (1993). Pitch accent in context: Predicting intonational prominence from text, *Artificial Intelligence* 63, 1-2

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer, *IEEE Trans. Speech and Audio Processing* 35, 3 (Mar.), 400-401

Khudanpur, S. and Wu, J. (2000). Maximum entropy techniques for exploiting syntactic, semantic and collocational dependencies in language modeling, *Computer Speech and Language* 14, 355-372

Kim, J. H. and Woodland, P. C. (2001). The use of prosody in a combined system for punctuation generation and speech recognition, *Proc. EUROSPEECH*

Kirchhoff, K., Bilmes, J., Das, S., Duta, N., Egan, M., Jin, G., He, F., Henderson, J., Liu, D., Noamany, M., Schone, P., Schwartz, R., and Vergyri, D. (2003). Novel approaches to Arabic speech recognition: report from the 2002 Johns-Hopkins summer workshop, *Proc. ICASSP*, Hong Kong, China

Kompe, R. (1997). Prosody in Speech Understanding Systems, Springer-Verlag

Lee, K. F. (1990). Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition, *IEEE Trans. Speech and Audio Processing* 38, 4 (Apr.), 599-609

Nakatani, C. H. and Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech, *J. Acoust. Soc. Am* 95, 3, 1603-1616

Ostendorf, M., Price, P. J., and Shattuck-Hufnagel, S. (1995). The Boston University Radio News Corpus, Linguistic Data Consortium

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., and Fong, C. (1991). The use of prosody in syntactic disambiguation, *J. Acoust. Soc. Am* 90, 6 (Dec.), 2956-2970

Rosenfeld, R. (2000). Two decades of statistical language modeling: where do we go from here? *Proceedings of the IEEE* 88, 8, 1270-1278

Shriberg, E., Stolcke, A., Hakkani-Tur, D., and Tur, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics, *Speech Communication* 32, 1-2 (Sep.), 127-154

Taylor, P., King, S., Isard, S., Wright, H., and Kowtko, J. (1997). Using intonation to constrain language models in speech recognition, *Proc. EUROSPEECH*

Waibel, A. (1988). Prosody and Speech Recognition, London: Pitman