IMPROVING THE ROBUSTNESS OF PROSODY DEPENDENT LANGUAGE MODELING BASED ON PROSODY SYNTAX DEPENDENCE

Ken Chen and Mark Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, Urbana, IL 61801
http://www.ifp.uiuc.edu/speech/

ABSTRACT

This paper presents a novel approach that improves the robustness of prosody dependent language modeling by leveraging the dependence between prosody and syntax. A prosody dependent language model describes the joint probability distribution of concurrent word and prosody sequences and can be used to provide prior language constraints in a prosody dependent speech recognizer. Robust Maximum Likelihood (ML) estimation of prosody dependent n-gram language models requires a large amount of prosodically transcribed data. In this paper, we show that the prosody-syntax dependence can be utilized to diminish the data sparseness introduced by prosody dependent modeling. Experiments on Radio News Corpus show that the prosody dependent language model estimated using our approach reduces the joint perplexity by up to 34% as compared with the standard ML-estimated prosody dependent language model; the word perplexity can be reduced by up to 84% as compared with the standard ML-estimated prosody independent language model. In recognition experiments, the language model estimated by our approach create an improvement of 1% in word recognition accuracy, 0.7% in accent recognition accuracy and 1.5% in intonational phrase boundary (IPB) recognition accuracy over a baseline prosody dependent language model.

1. INTRODUCTION

Prosody refers to the suprasegmental features of natural speech (such as rhythm and intonation) that are used to convey 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 [1]. In automatic speech understanding systems, prosody has been previously used to disambiguate syntactically distinct sentences with identical phoneme strings [2], infer punctuation of a recognized text [3], segment speech into sentences and topics [4], recognize the dialog act labels [5], and detect speech disfluencies [6]. 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. [7] 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:

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

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.

Systems described by equation (1) is similar in appearance to a system proposed by Heeman [8] in which parts-of-speech (POS) sequences P' rather than prosody sequences P are modeled jointly with the word sequences W. Heeman has shown that the accuracy of the language model $p(W,P')$ can be improved significantly by modeling the inter-dependence between W and P' using decision trees. However, the acoustic model in Heeman's system is not dependent on P': $p(O|W,P') \approx p(O|W)$, due to the absence of a clear relationship between acoustic signal and POS. Heeman did not provide any word recognition results using his system and primarily used his system to recognize POS, discourse markers, speech repairs and intonational phrase boundaries by modeling in language model the inter-dependence among these variables.

The prosody dependent recognition system we propose in (1) have the advantage that both the acoustic model and the language model can be potentially improved through their dependence on prosody: the prosody induced acoustic variation can be modeled in the prosody dependent acoustic model, and the inter-dependence between concurrent word and prosody sequence can be modeled in the prosody dependent language model. Chen et al. [9] have shown using an information-theoretic analysis that it is possible for a prosody-dependent speech recognizer to result in improved 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 Radio News Corpus [10], 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.

In reference[9], 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 $p(W, P)$, which 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 et al. [11] 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. Arnfield [12] gives an example in his dissertation: the word "witch" and "which", having identical acoustic observation, 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. [9], a prosody dependent language model can significantly improve word recognition accuracy over a prosody independent language model, given the same acoustic model.

Similar to prosody independent language modeling, n-gram models can be conveniently used for the 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 $p(w_j, p_j|w_i, p_i)$ (the probability of observing token $[w_j, p_j]$ given token $[w_i, p_i]$) can be estimated using the following equation:

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

where $n(\cdot)$ is the number of the n-grams observed in the training set. Equation (2) treats each prosody dependent word token $[w, p]$ as a distinct unit, resulting in a recognizer that has $|p|$ times larger vocabulary size than does a standard prosody independent recognizer (the $|p| = 1$ case). 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 paper, we propose to improve the robustness of prosody dependent language modeling by utilizing the dependence between prosody and syntax. There are evidences indicating that syntax is a strong confining factor for prosody. For example, conjunctions (e.g., "but", "so"), having high probability of occurring at phrase initial positions, can never appear at phrase final positions; 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 [12] 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). Our preliminary experiment on Radio New Corpus also indicates that parts of speech can predict the presence of pitch accent with an accuracy of around 85%. Motivated by these results, we propose to use parts of speech as word classes to facilitate the estimation of prosody dependent n-grams. In section 2, we propose an approach that increases the robustness of prosody dependent n-gram modeling by leveraging the prosody-syntax dependence. Section 3 reports experiments and results on the Radio News Corpus. Section 4 gives the conclusion.

2. THE METHOD

In this section, we propose an approach that improves the robustness of prosody dependent n-gram language modeling 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 language models.

2.1. Class-Dependent Prosodic Language Model

The conditional 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 $p(w_j|w_i)$ using following equation:

$$\begin{aligned} & p(w_j|w_i,p_i) \\ &= \frac{p(p_i,w_j|w_i)}{p(p_i|w_i)} \\ &= \frac{p(p_i|w_j,w_i)p(w_j|w_i)}{p(p_i|w_i)} \\ &\approx \frac{\sum_{c_i,c_j} p(p_i|c_i,c_j)p(c_i,c_j|w_i,w_j)p(w_j|w_i)}{\sum_{w_j}\sum_{c_i,c_j} p(p_i|c_i,c_j)p(c_i,c_j|w_i,w_j)p(w_j|w_i)} \end{aligned} \quad (3)$$

where c_i and c_j are the word classes of w_i and w_j respectively. Parts of speech (POS), representing the syntactical role of word, is chosen as word class in this research due to the known strong dependence between prosody and syntax [12].

The approximation in equation (3) assumes that p_i (the prosody on the previous word) is dependent on the POS context but independent of the actual word context:

$$p(p_i|c_i,c_j) \approx p(p_i|c_i,c_j,w_i,w_j). \quad (4)$$

Similarly, 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 probability $p(w_j|w_i,p_i)$:

$$\begin{aligned} & p(w_j,p_j|w_i,p_i) \\ &= p(p_j|w_j,w_i,p_i)p(w_j|w_i,p_i) \\ &= \sum_{c_i,c_j} p(p_j|c_j,c_i,w_j,w_i,p_i)p(c_j,c_i|w_j,w_i,p_i)p(w_j|w_i,p_i) \\ &\approx \sum_{c_i,c_j} p(p_j|c_j,c_i,p_i)p(c_j,c_i|w_j,w_i)p(w_j|w_i,p_i). \end{aligned} \quad (5)$$

Following assumptions are required in deriving equation (5):

$$p(p_j|c_j,c_i,p_i) \approx p(p_j|c_j,c_i,w_j,w_i,p_i), \quad (6)$$

and

$$p(c_j,c_i|w_j,w_i) \approx p(c_j,c_i|w_j,w_i,p_i). \quad (7)$$

Equation (6) again 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 (7) 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. In this paper, we assume that the use of prosody to resolve POS ambiguity is statistically rare in our corpus.

Together, equations (3) and (5) provide an approach to calculate the prosody dependent bigram probability $p(w_j,p_j|w_i,p_i)$ based on the regular prosody independent bigram probability $p(w_j|w_i)$ and three additional probability mass functions: $p(p_i|c_i,c_j)$, $p(p_j|c_j,c_i,p_i)$, and $p(c_j,c_i|w_j,w_i)$. $p(c_j,c_i|w_j,w_i)$ describes the stochastic mapping between a word pair and the associated POS pair. In most cases, it is deterministic (constantly equals to 1) in the sense that a word pair can only be associated with one 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 $p(p_i|c_i,c_j)$ and $p(p_j|c_j,c_i,p_i)$ describe the inter-dependence between prosody and parts-of-speech, and can be very robustly estimated from a small database due to the limited variety of POS tokens and prosody tokens. Note that equations (5) is possibly more accurate than equation (3) because the approximations are made only in numerator while equation (3) has approximations in both numerator and denominator.

2.2. Methods for Smoothing the Language Models

Two popular techniques can be used to smooth the resulting language model: the backoff scheme and the 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 [13]:

$$p_b(w_j,p_j|w_i,p_i) = \begin{cases} d_r p(w_j,p_j|w_i,p_i), & \text{if exists} \\ b(w_i,p_i)p(w_j,p_j), & \text{else} \end{cases} \quad (8)$$

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 to 1, i.e.,

$$b(w_i, p_i) = \frac{1 - \sum_{j \in B} p(w_j, p_j | w_i, p_i)}{1 - \sum_{j \in B} p(p_j | w_j)}, \quad (9)$$

where B is the set of all prosody dependent word labels $[w_j, p_j]$ whose bigram probabilities can be calculated from equation (3) and (5).

The bigram probabilities calculated from equation (3) and (5) can be interpolated with the bigram probabilities estimated directly from the data (equation (2)). Let p_c be the probabilities calculated by equation (3) and (5), and p_m the probabilities estimated by equation (2), the interpolated probability p_i can be obtained using:

$$\begin{aligned} & p_i(w_j, p_j | w_i, p_i) \\ = \; & \lambda p_c(w_j, p_j | w_i, p_i) + (1 - \lambda) p_m(w_j, p_j | w_i, p_i), \end{aligned} \quad (10)$$

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. EXPERIMENTS AND RESULTS

3.1. 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 [10]. 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. For the labnews stories (a subset of the Radio New Corpus recorded without noise in a phonetics laboratory), it is found that 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 [14]. 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, representing high tone, low tone, and high tone followed by pitch downstep, respectively. 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 [10]. Presence versus absence of pitch accent is transcribed with 91% inter-transcriber agreement.

In the experiments we reported in this paper, the original ToBI labels are simplified: accents are only distinguished by presence versus absence, word boundaries are only distinguished by intonational phrase boundary versus normal word boundary. 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").

3.2. 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 (5) only, and a model LPDC2 computed using both equation (3) and equation (5). The difference between LPDC1 and LPDC2 is that in LPDC1, the intermediate prosody dependent bigram probabilities $p(w_j | w_i, p_i)$ required by equation (5) are estimated directly from data using their relative frequency count; whereas in LPDC2 they are computed from the prosody independent bigram probabilities $p(w_j | w_i)$ using equation (3). The models LPDC1 and LPDC2 were linearly interpolated with LPDM using equation (10),

	LPI	LPDM	LPDC1	LPDC2
Joint Perp.		354	282	235
Word Perp.	130	43	30	21
Unseen bigrams	931	1255	1108	953
Total bigrams	12100	14648	37341	81431

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).

with interpolation weights λ optimized over the development-test set. Table 1 lists the results of this experiment.

Compare the performance among the prosody dependent language models: LPDM, LPDC1 and LPDC2. Both LPDC1 and LPDC2 have much smaller joint perplexity (the perplexity measured using prosody-dependent word tokens) than LPDM: the perplexity of LPDC1 is 24% less than that of LPDM, while that of LPDC2 is 34% smaller. Class-dependent modeling increases the number of bigrams whose probabilities can be estimated without backoff: 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 marginalized the joint perplexity over all possible prosody sequences and obtained the word perplexity. As can be seen in the third row of Table 1, word perplexity of LPDC2 is reduced by as large as 84% from that of LPI. Note that the word perplexities of prosody dependent language models are only weakly comparable with that of the prosody independent language models in predicting word recognition performance. The word recognition power of the prosodic dependent language model is only prominent when it is coupled with an effective prosody dependent acoustic model.

3.3. Word Recognition

Encouraged by the great reduction in perplexity, we conducted word and prosody recognition experiment 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

	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.

in APD contains 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 [9]. API contains monophone models adopted from the standard SPHINX set [15] 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: the 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 plus LPDC1, and a prosody dependent recognizer RDC2 using APD plus 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.

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 benefitting 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 1). It is possible that the approximations in equation (3) do not accurately represent the probabilities of all of these bigrams, and that therefore the increased flexibility harms word recognition accuracy.

4. CONCLUSION

In this paper, 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 prosody dependent language model estimated using our proposed approach has achieved as much as 34% 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 paper shows that prosody-syntax dependence can be used to reduce the uncertainty in modeling concurrent word-prosody sequences.

5. REFERENCES

[1] L. Hahn, "Native speakers' reactions to non-native stress in English discourse," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1999.

[2] P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *J. Acoust. Soc. Am.*, vol. 90, no. 6, pp. 2956-2970, Dec. 1991.

[3] J. H. Kim and P. C. Woodland, "The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition," in *Proc. EUROSPEECH'01*.

[4] E. Shriberg, A. Stolcke, D. Hakkani-Tur and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communication*, vol. 32, issues 1-2, pp. 127-154, Sep. 2000

[5] P. Taylor, S. King, S. Isard, H. Wright and J. Kowtko, "Using intonation to constrain language models in speech recognition," in *Proc. EUROSPEECH'97*.

[6] C. H. Nakatani and J. Hirschberg, "A corpus-based study of repair cues in spontaneous speech," *J. Acoust. Soc. Am.*, vol. 95, no. 3, pp. 1603-1616, 1994.

[7] K. Chen, S. Borys, M. Hasegawa-Johnson and J. Cole, "Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries," in *Proc. EUROSPEECH'03*.

[8] P. Heeman and J. Allen, "Speech repairs, intonational phrases and discourse markers: modeling speakers' utterances in spoken dialog," *Computational Linguistics*, vol. 25, no. 4, 1999.

[9] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S. Kim, J. Cole and J. Choi, "Prosody dependent speech recognition on Radio News," in review.

[10] M. Ostendorf, P. J. Price and S. Shattuck-Hufnagel, *The Boston University Radio News Corpus.* Linguistic Data Consortium, 1995.

[11] R. Kompe, "Prosody in speech understanding systems," *Lect. Notes in Artificial Intelligence*, Springer-Verleg, 1307:1-357, 1997.

[12] S. Arnfield, "Prosody and syntax in corpus based analysis of spoken English," Ph.D. dissertation, University of Leeds, Dec. 1994.

[13] S. M. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 35, no. 3, pp. 400-401, March 1987.

[14] M. E. Beckman and G. A. Elam, "Guidelines for ToBI labelling," 1994, http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html.

[15] K. F. Lee, "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 4, pp. 599-609, April 1990.