An Intonational Phrase Boundary and Pitch Accent Dependent Speech Recognizer

Ken Chen[1], *Mark Hasegawa-Johnson*[1] *and Sung-Suk Kim*[2]

1. ECE Department, University of Illinois at Urbana-Champaign
Urbana, IL 61801, U. S. A.
http://www.ifp.uiuc.edu/speech/
2. School of Computer and Information, Yong-In University, South Korea

ABSTRACT

Does prosody help word recognition? In this paper, we propose a novel probabilistic framework in which word and phoneme are dependent on prosody in a way that improves word recognition. We describe the idea of prosody dependent speech recognition by building a prosody dependent speech recognizer that conditions word and phoneme models on two important prosodic variables: intonational phrase boundary and pitch accent. It is known that intonational phrase boundaries induce salient lengthening to the phrase-final speech units, while pitch accents induce distinct pitch variation on the accented syllables. Effective prosody-discriminative hidden Markov models (HMMs) can be built by conditioning on prosody only a small subset of HMM distributions: the duration PDFs and the acoustic-prosodic observation PDFs. The prosody dependence of the acoustic-phonetic observation PDFs is ignored in our investigation, resulting in a prosody dependent recognizer that has a good trade-off between performance and parameterization. To accurately model the duration of the prosody dependent allophonic models, explicit duration hidden Markov model (EDHMM) is used for both training and decoding. A new acoustic-prosodic feature, transformed from the normalized F0 contour by an artificial neural network (ANN), is incorporated into the acoustic feature vector for the acoustic modeling of accent induced pitch variation. This prosody dependent speech recognizer is able to improve word recognition accuracy by an absolute 1.8% over prosody independent recognizers on the Boston University Radio News Corpus which is prosodically transcribed using ToBI labeling system.

Keywords: prosody, duration, pitch, HMM, ANN, word recognition accuracy, ToBI.

1. INTRODUCTION

Does prosody help word recognition? 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]. For automatic Large Vocabulary Continuous Speech Recognition (LVCSR), there is no straightforward answer.

Prosody is potentially useful in automatic speech understanding systems for at least four reasons. First, prosody correlates with syntax: Price et al. [2] showed that prosody may be used to disambiguate syntactically distinct sentences with identical phoneme strings, while Kim et al. [3] have demonstrated that prosody may be used to infer punctuation of a recognized text. Second, prosody correlates with meaning: for example, Taylor et al. [4] have used prosody for the purpose of recognizing the dialog act labels of utterances. Third, prosody is useful for the detection and subsequent processing of speech disfluencies [5]. Finally, prosody may be useful as prior conditioning information for the correct phoneme labeling of an ambiguous acoustic signal.

In an automatic speech recognition system, prosody may be recognized before, after, or simultaneously with the recognition of phonemes and words. The ordering of the word-recognition module and the prosody-recognition module depends on the intended purpose of prosody recognition. Systems that intend to use prosody only for the purpose of semantic, syntactic, or disfluency processing often implement a prosodic post-processing strategy, in which the input to the prosody recognizer includes a time-aligned word graph generated by an initial prosody-independent speech recognizer. The advantage of a post-processor strategy is greater accuracy, won by the of syllable-timed acoustic features (e.g., average F0 during the syllable of interest) and word string information [6] [7]. These advantages are compelling in many applications: all reported uses of prosody in commercial speech understanding systems use a post-processor model of prosody recognition. The disadvantage of a post-processor strategy is that the front-end recognizer is unable to use prosody to aid in the phonetic labeling of ambiguous acoustic signals. Kompe [6] demonstrates both theoretically and empirically that a prosody post-processor can improve the search time of a speech recognizer, but never its word recognition accuracy.

In this paper, we present a novel probabilistic framework in which prosody and word are modeled jointly and recognized simultaneously. In section 2, we discuss prosody induced allophonic variation reported by phoneticians and computational linguists, and briefly introduce the Boston University Radio News Corpus and the ToBI labeling system that are used in this research. In section 3, we present the mathematical framework in which word and prosody are modeled as joint events through prosody dependent allophonic models. Specifically, we propose to condition word and phoneme models on two important prosodic variables: intonational phrase boundary and pitch accent. In this recognizer, good tradeoff between performance and parameterization can be achieved by restricting the prosody dependence on a small subset of distributions: the duration PDFs and the acoustic-prosodic observation PDFs. Section 4 discusses the labeling system with which prosody dependent speech recognition can be implemented in a conventional speech recognizer. Section 5 introduces our approach of creating an acoustic-prosodic feature for the modeling of accent induced pitch variation. Section 6 reports the recognition experiments and results,

and conclusions are given in section 7.

2. BACKGROUND

Prosodically Induced Allophonic Variation

Prosody refers to the suprasegmental features of natural speech, such as rhythm and intonation. Native speakers use prosody to convey paralinguistic information such as emphasis, intention, attitude and emotion. The prosody of a word sequence is determined by the values of a set of prosodic variables such as prosodic phrase boundary, pitch accent, lexical stress, syllable position in word and disfluency, etc. Among these prosodic variables, pitch accent and intonational phrase boundary are the most important ones that have salient acoustic correlates. A pitch accent is an unusually high F0 (possibly a local maximum) or an unusually low F0 (possibly a local minimum) designed to draw attention to the important word [8]. The presence of a pitch accent correlates with other changes in the acoustic signal: accented vowels tend to be longer and less subject to coarticulatory variation [9], while accented consonants are produced with greater closure duration [10], greater linguopalatal contact [11], longer voice onset time, and greater burst amplitude [12]. Knowledge of pitch accent placement would therefore be useful prior information for accurate acoustic modeling. Intonational phrase boundaries, which segments an utterance into intonational phrases, not only introduce distinct pitch contour (boundary tones) on the phrase final speech segments, but also affect the acoustic realization of neighboring phonemes: phonemes preceding phrase boundaries are lengthened considerably [13] and consistently [14], phonemes both preceding and succeeding intonational phrase boundaries have more extreme lingual articulations [11]. These acoustic correlates of prosody can be modeled in HMM-based automatic speech recognizers through prosody dependent allophone models that are created by splitting prosody independent monophone or triphone models with respect to certain prosody variables. The allophonic variation induced by pitch accent and intonational phrase boundary potentially affects HMMs in their PDFs of duration, acoustic-prosodic observation (e.g., pitch and energy) and acoustic-phonetic observation (e.g., subband energy or cepstral coefficients).

The Radio News Corpus

To train automatic speech recognizers that are aware of the prosody induced allophonic variation, a large scale prosodically labeled speech database is required. The Boston University Radio News Corpus is one of a few such corpora that is designed for study of prosody. 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, hence provide speech with a *natural but controlled* style [15], 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 prosodic labeling system represents prosodic phrasing, phrasal prominence and boundary tones, using the ToBI system for American English [16]. In ToBI, break indices are marked to indicate the degree of decoupling between each pair of words. The intonational phrase boundaries are marked by break index of 4, and is most reliably labeled due to the existence of boundary tones. Seven types of accent tones are labeled: H*, !H*, L+H*, L+!H*, L*, L*+H and H+!H*, where H and L correspond to high and low targets, "!" indicates downstep and the asterisk indicates tone alignment. Intermediate phrase boundaries are marked with three tones (L-, !H- and H-), and intonational phrase boundaries are marked with one of these tones plus a phrase boundary tone (L% or H%). The ToBI system has the advantage that it can be used consistently by labelers for a variety of styles. In this corpus, there is 91% agreement on presense versus absence of pitch accents and 95% agreement for the five break index levels within certain uncertainty level among two labelers with different labeling styles.

3. PROSODY DEPENDENT SPEECH RECOGNITION

In this section, we describe the mathematical framework for prosody dependent word and phoneme modeling. The task of prosody dependent speech recognition, given a sequence of observed short-time vectors $O = (o_1, ...o_T)$ of the acoustic features, is to find the sequence of word labels $W = (w_1, ..., w_M)$ and the sequence of prosody labels $P = (p_1, ..., p_M)$ that maximizes the recognition probability:

$$[\hat{W}, \hat{P}] = \arg\max p(O, Q, W, P), \quad (1)$$

where $Q = (q_1, ..., q_L)$ is a sequence of sub-word units, typically allophones dependent on phonetic context. Ostendorf et al. [17] suggested expanding Eq. (1) as:

$$[\hat{W}, \hat{P}] = \arg\max p(O|Q, H)p(Q, H|W, P)p(W, P), \quad (2)$$

where $H = (h_1, ..., h_L)$ is a sequence of discrete "hidden mode" vectors describing the prosodic states of Q, $p(O|Q, H)$ 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. Eq. (2) proposes that every distinct combination of the state variables q and h should be modeled using a distinct acoustic model. In the most straightforward implementation of Eq. (2), a recognizer aware of $|h|$ different prosodic contexts would require $|h|$ times as many trainable parameters as a prosody-independent recognizer. In our experiments, we find that the number of parameters required to directly implement Eq. (2) is rarely justified by a proportional increase in recognition accuracy. We therefore propose to model only the most salient and widely reported acoustic effects of prosody: phrase-final duration lengthening and accent induced pitch variation, but ignore prosody induced spectral variation.

In our previous work [18], we investigated the duration lengthening of speech segments in the vicinity of intonational phrase boundaries and showed that the phrase final lengthening can be reliably modeled to improve both the word and boundary recognition accuracy. In this paper, we extended the prosody dependence to another common prosody attribute: pitch accent, by making h a two dimensional prosodic vector: $h = [a, b]$, where a is a binary variable indicating if a phoneme q is accented, and b is another binary variable indicating if q is lengthened for being in the phrase final position. The observation vector O is augmented to include two independent feature streams X and Y: $O = [X, Y]$, where X is a conventional acoustic-phonetic observation vector (typically cepstral coefficients) and Y is an acoustic-prosodic observation vector (a transformation of normalized pitch). The phrase-final lengthening is modeled by conditioning the state residency time or "duration" of a phonetic

state of a HMM on the prosodic variable b. The accent induced pitch variation is modeled by conditioning the acoustic-prosodic observation Y on both the phoneme q and the prosodic variable a. We propose to use prosody-dependent explicit duration hidden Markov models (EDHMM) in order to precisely model the intonational phrase boundary induced duration lengthening. The EDHMM of phoneme q_i under prosodic state b_i consists of a sequence of hidden state variables $S_i = (s_{i1}, ..., s_{iN})$, each of which persists for duration d_{ij}, and each of which produces a length-d_{ij} sequence of observation vectors denoted O_{ij}. If, as we propose, the prosodic variables influence only phoneme duration and pitch, then the probability of observing matrix $O_i = [O_{i1}, \ldots, O_{iN}]$ is

$$\begin{aligned} & p(O_i|q_i, h_i) \\ &= p(X_i, Y_i|S_i, h_i)p(S_i|q_i, h_i) \\ &= \prod_{j=1}^{N} p(X_{ij}, Y_{ij}|s_{ij}, a_i, b_i)p(d_{ij}|s_{ij}, b_i)p(S_i|q_i) \\ &= \prod_{j=1}^{N} p(X_{ij}|s_{ij})p(Y_{ij}|s_{ij}, a_i)p(d_{ij}|s_{ij}, b_i)p(S_i|q_i). \end{aligned} \quad (3)$$

In this paper, the word-level prosody variable p_m in Eq. (2) takes eight possible values indicating the relative position of the word w_m in an intonational phrase (phrase initial, phrase medial, phrase final, or a one-word intonational phrase) and its binary prominence level (accented, unaccented). The phone-level prosodic variable b_i takes only two possible values, indicating whether the corresponding allophone q_i is phrase final, where the phrase final phonemes are defined as the vowel nuclei and any coda consonants in the intonational phrase final syllables. The prosodic variable a_i is also binary, indicating whether the corresponding allophone q_i receives a pitch accent. Under these definitions, we achieve a four-way prosodic distinction for each phonetic state s_{ij}: neutral (default), accented, lengthened, accented+lengthened. Note that the number of parameters used to model s_{ij} is not increased by four times because all four allophonic variants of s_{ij} share the same acoustic-phonetic observation PDF $p(X_{ij}|s_{ij})$. Meanwhile, the duration PDFs of the allophonic variants having the same level of lengthening (same value of b) are shared, and so do the acoustic-prosodic observation PDFs of those having the same level of prominence (same value of a).

The pronunciation model $p(Q, H|W, P)$ is implemented using a prosody-dependent dictionary in which the connection between the word level prosody variable p_m and the phoneme level prosody variables a_i and b_i are explicitly encoded. The language model $p(W, P)$ is implemented as a prosody-dependent bigram:

$$p(W, P) = p(w_1, p_1) \prod_{m=2}^{M} p(w_m, p_m|w_{m-1}, p_{m-1}). \quad (4)$$

4. THE LABELING SYSTEM FOR PROSODY DEPENDENT SPEECH RECOGNITION

Prosody is incorporated into automatic speech recognition by first creating prosody-tagged word transcriptions using the prosody labels in the break files and the tone files in the Radio News Corpus. In the break files, transcriptions of ToBI break indices are given at each word boundary. The intonational phrases boundaries (break level 4 and above) can be found by time-aligning the break transcription in the break file with the corresponding word transcription. Symbol b4 is used as prefix or postfix to tag the words to indicate their positions in an intonational phrase. Specifically, a word W is labeled as W_b4, b4_W or b4_W_b4 if it is phrase final, phrase initial, or a one-word intonational phrase respectively. Similarly, a word receives a "!" tag if any pitch accent labeled in the tone file falls within that word interval. The boundary and accent tags are cumulative, creating maximally an eight way distinction for each word in the vocabulary indicating its 4 possible relative positions: phrase initial, phrase medial, phrase final, or a one-word intonational phrase, and its binary prominence level: accented, unaccented.

The prosody dependence is propagated from word level to the phonetic level through prosody dependent dictionaries. It is possible to create different prosody dependent phoneme transcriptions from the same prosody dependent word transcription depending on how many prosodic distinction one intends to make at the phonetic level. For example, to model the phrase-final lengthening effects, the phonetic labels corresponding to the final vowel (FV) and final coda consonants (FC) in a phrase final word W_b4 or b4_W_b4 are appended with the _b4 postfix while other phonetic labels are unchanged such that the phonetic models of phrase final phonemes can be distinctively modeled from other non-phrase-final phonemes. To mark pitch accent at phonetic level, we assume that only the phonemes in the primary lexical-stressed syllable of an accented word are accented, and ignore the possibility that sometimes accents can appear over consecutive syllables (e.g., emphatic accents) or realize on in a non-primary-stressed syllables (e.g., contrastive accents). Using this labeling scheme, we are able to create prosody dependent transcriptions at both word and phonetic levels, with prosody dependence selectively tagged to specific words and phonemes depending on the specific prosody effects that we want to model in our recognizer.

5. THE ACOUSTIC-PROSODIC FEATURE

To build accent dependent allophonic models, an acoustic-prosodic feature representing the probability of pitch accent at each frame is incorporated into the observation feature stream. In this study, we extracted the fundamental frequency $f_0(t)$ from speech using the formant program in Entropic XWAVE. The probability of voicing (PV) is output simultaneously as a confidence measure to the extracted $f_0(t)$. This raw $f_0(t)$ usually contains a small amount of pitch doubling and halving errors that are inevitable for any existing type of pitch trackers. To avoid these pitch doubling and halving errors, we trained a 3 mixture Gaussian classifier whose means of the mixture components are restricted to $1/2\mu$, μ, and 2μ respectively, where μ is the mean f_0 estimated from the whole utterance read by a single speaker, and eliminated those f_0 that have higher probability of belonging to either the doubling cluster or halving cluster than to the center cluster. We then divide $f_0(t)$ by μ and convert it to log scale to remove some amount of speaker dependence:

$$\hat{f}_0(t) = \log(f_0(t)/\mu + 1). \quad (5)$$

The $f_0(t)$ with small PVs are normally extracted from the non-vocalic frames and are not reliable. Thus, we eliminate those $\hat{f}_0(t)$ whose PV are smaller than a heuristic threshold and replaced them by the linear interpolated values $\tilde{f}_0(t)$ based on the $\hat{f}_0(t)$ that have PVs greater than the threshold. The linear inter-

polation of f_0 has been used by Kompe [6] and has been shown to be a good approximation.

The f_0 normalized above does not necessarily has a Gaussian distribution. To use it as an acoustic feature for HMM, we need to transform it to a new variable Y whose distribution can be accurately modeled as a mixture Gaussian PDF. Artificial Neural Networks (ANN) suits ideally for this task because of its ability to approximate any feature transformation function given enough training data. This feature transformation problem is equivalent to a pitch accent recognition problem where ANN is used to compute the a posteriori probability of pitch accents at each frame. This a posteriori probability, if accurately estimated, is approximately Gaussian distributed over accent/unaccent events, and can be modeled by mixture-Gaussian models in HMM.

In ToBI, three fundamental pitch movements: high (H), low (L), and downstepped (!H) have been posited. The vast majority of pitch accents in the Radio News corpus are centered on a high pitch movement (71%) or a downstepped pitch movement (25%). Only about 5% of the accents are transcribed as L*. Dainora [19] argues that !H and H movements are not linguistically distinct and should therefore not be distinctly recognized. The nonuniform prior distribution of accent classes posit a problem to the ANN classifier which usually desires to have approximately equal number of examples for each class. To circumvent these problems, we grouped H* and !H* accents into a single accent class denoted by a single letter: A, and grouped L* and non-accented region (including boundary tones) into an unaccented label: UA. The task of ANN is to classify the pitch contour into consecutive accent and unaccent regions. At prosodic phrase boundaries, there are boundary tones (for example, L-H% and H-L%) that have pitch patterns similar to those of pitch accents but have completely different linguistic meaning. It is unclear whether boundary tones should be grouped with pitch accents in this recognition problem. It is likely that the pitch movement of boundary tones can cause some confusion to the pitch accent classifier. These problems are not investigated in this paper but will be investigated in our future work.

Different types of ANNs are attempted that model pitch accents as either static patterns or dynamic patterns. For the case of static pattern recognition, a multi-layer perceptron (MLP) $g(\cdot)$ is trained under the objective of minimizing the mean square error between the MLP output and a teaching signal that is designed based on the pitch accent labels. About 85% pitch accent event recognition rate is achieved by using MLP. Pitch accents can also be modeled as dynamic patterns. Kim et al. [20] used a time-delayed recurrent neural network (TDRNN) to recognize pitch accents, and around 90% recognition rate is achieved. After the ANN is trained, the acoustic-prosodic observation $Y(t)$ can be generated using:

$$Y(t) = g(\tilde{f}_0(t)). \quad (6)$$

6. EXPERIMENTS AND RESULTS

In this section, we report the recognition results for the prosody dependent recognition system that we proposed in section 3.

HMMs and Features

In all experiments, a 3-state HMM with no skip is used to model all the prosody-dependent allophones. The acoustic-phonetic observation PDF $p(X_{ij}|s_{ij})$ in Eq. (3) is modeled as 3-component mixture Gaussians, and the acoustic-prosodic observation PDF $p(Y_{ij}|s_{ij}, a_i)$ is modeled as a single Gaussian. The baseline prosody-independent phoneme set is adopted from the SPHINX phoneme set [21] after eliminating some of the low-frequency function-word-dependent phonemes. A 32 dimensional feature vector consists of 15 MFCC coefficients, energy, and their deltas are used as acoustic-phonetic features. A one dimensional acoustic-prosodic feature, as has been discussed in section 5, is included in the acoustic feature stream.

The prosodically labeled data consist of 300 utterances (about 3 hours of speech sampled at 16Khz) read by five professional announcers (3 female, 2 male) consisting of a vocabulary of 3777 words. Training and test sets are formed by randomly selecting 90% of the utterances for training and the rest 10% for testing.

	HMM	**EDHMM**
Phone Corr.(%)	64.82	64.84
Phone Acc.(%)	50.98	51.86

Table 1: Phoneme Recognition experiments on TIMIT.

	HMM			**EDHMM**		
	Corr	Acc	#para	Corr	Acc	#para
PI	14.05	2.38	39000	14.32	2.68	43414
PD	33.74	18.9	39789	33.76	19.62	47053

Table 2: % Allophone recognition correctness and accuracy on prosody dependent allophones, and number of parameters of the allophone models. Both PI and PD contain 166 allophones.

Results and Discussion

Three major recognition experiments were conducted: a prosody-independent phoneme recognition experiment using the TIMIT database (Table 1), a prosody-dependent phoneme recognition experiment using the Radio News Corpus (Table 2) and a prosody-dependent word and prosody recognition experiment using the Radio News Corpus (Table 3).

To compare the performance of EDHMM with standard HMM, we conducted phoneme recognition experiments on the TIMIT database using standard 48 phonemes modeled by HMMs of 3 non-skipping states and 3 mixture Gaussians per state. The phoneme recognition accuracy under no grammar condition is improved by .9%, as shown in Table 1.

To precisely measure the influence of prosodic context on the prosody dependent allophonic models, we conducted prosody-dependent allophone recognition experiments on the Radio News Corpus. Two sets of allophone models were constructed: a prosody-dependent set PD whose prosodic contexts are differentiated by the duration PDFs and the acoustic-prosodic observation PDFs, as shown in Eq. (3), and a baseline prosody-independent set PI whose prosodic contexts are logically distinct but physically the same, i.e., the duration PDFs under different prosodic contexts are tied and the acoustic-prosodic observation PDFs are removed. By comparing the prosody-dependent allophone recognition correctness and accuracy of PD and PI models with a null grammar (every allophone sequence equally likely), it is possible to assess the strength of the dependence of the allophonic models over the prosodic context in the Radio News Corpus. Table 2 shows the results of this experiment. As can be seen in Table 2, the allophone recognition accuracies of PD significantly exceed those of PI and the total numbers of parameters in PD are only

	AM	LM	HMM	EDHMM
Word	PI	PI	74.89	75.15
	PI	PD	75.52	75.67
	PD	PD	76.50	76.62
Accent	PI	PI	55.41	55.37
	PI	PD	76.75	76.92
	PD	PD	79.61	79.65
IPB	PI	PI	84.47	84.43
	PI	PD	85.33	85.53
	PD	PD	85.49	85.65

Table 3: % word, accent and intonational phrase boundary (IPB) recognition accuracy using PI and PD acoustic models in combination with PI and PD language models.

slightly greater than those in PI. This indicates that prosody is both effectively and efficiently modeled in the allophone models.

To measure the overall performance of prosody dependent recognition, we conducted word recognition experiments and prosody recognition experiments using two types of Acoustic Models (AM) and two types of bigram Language Models (LM). The two types of acoustic models are PI and PD which have been used in the above prosody dependent allophone recognition experiment. The two types of language models are denoted as PI and PD as well. Here, PI denotes a LM that contains only plain words with no prosody tags; and PD is a LM that has the maximal prosody dependence in which each word can have at most eight prosody dependent variants. We found that the entropy of the test text decreases from 2.05 bits in PI to 1.72 bits in PD, at a cost of increasing the number of parameters in the language models from 5380 to 14751 due to the increasing of the size of the vocabulary. This result can be explained by the strong correlation between prosody and word strings in Radio News. By construction, this database includes many word string repetitions. Word strings in the training text often re-appear in the test text with the same prosodic pattern. Hence, in this case, prosody helps increase the predictability of test text and thus reduce its entropy. This improvement of language modeling is evident by comparing the PI+PI results with the PI+PD results in Table 3, where it shows that with the same acoustic model PI, the language model PD can improve word recognition accuracy by about .6% over the language model PI. After switching to acoustic model PD, the word recognition accuracy can be further improved because the interaction between the prosody-dependent acoustic model and prosody-dependent language model increases the likelihood of the word hypotheses that are uttered with regular prosody, and reduces the likelihood of the word hypotheses uttered with irregular prosody. This statement is supported by Table 3 which shows that the word recognition accuracy of PD+PD+EDHMM has improved about 1.8% over the baseline system PI+PI+HMM.

Unlike the conventional prosody independent recognizer, the prosody dependent recognizer not only output word transcriptions but also output prosody transcriptions. Table 3 also reports the accent and boundary recognition accuracy from this prosody-dependent recognition experiment. In this case, we compared the decoded accent and boundary transcriptions with the reference accent and boundary transcriptions. The recognition accuracy of pitch accent increases from the chance 55.41% to 79.65% with prosody dependence modeled. Note that this accent recognition results are not directly comparable to the accent event recognition results of ANN in section 5, because it is dependent on the word recognition accuracy. The recognition accuracy of intonational phrase boundary has improved by only 1.1% because of the increased number of word boundary insertions. Roughly 15% of the word boundaries in this database are intonational phrase boundaries, thus simply setting all word boundaries to be b0 gives a boundary recognition correctness of about 84.5%.

7. CONCLUSIONS

In this paper, a prosody dependent speech recognizer that models word and prosody in a unified probabilistic framework is proposed. We modeled the prosody induced allophonic variation by modeling the prosody dependence of HMM state duration PDFs and the acoustic-prosodic observation PDFs while ignoring the prosody dependence of the acoustic-phonetic observation PDFs. By so doing, we achieved a good tradeoff between performance and parameterization. To accurately model the phoneme duration, we implemented the recognizer using the explicit duration hidden Markov model (EDHMM) and compared it with systems using standard HMM. To model pitch accent, a new acoustic prosodic feature, transformed by an ANN from normalized pitch, is incorporated into the acoustic observation, and is modeled by a single Gaussian PDF. In the word and prosody recognition experiment on the Radio News corpus, we found that the proposed prosody dependent recognizer improves word recognition accuracy over prosody independent recognizers by as large as 1.8% using EDHMMs. The best tradeoff between performance and parameterization is achieved by a prosody-dependent HMM recognizer which increases word recognition accuracy by 1.6% with only 1.7% parameter increase in the acoustic model and about 21% parameter increase in the language model.

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

[5] 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.

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

[7] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. on Speech and Audio Processing*, vol. 2, No. 4, pp. 469-481, Oct. 1994.

[8] M. E. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook*, vol. 3, pp. 255-309, 1986.

[9] T. Cho, "Effects of prosody on articulation in English," Ph.D. dissertation, UCLA, 2001.

[10] K. DeJong, "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 369-382, 1995.

[11] C. Fougeron and P. A. Keating, "Articulatory strengthening at edges of prosodic domains," *J. Acoust. Soc. Am.*, vol. 101, no. 6, pp. 3728-3740, June 1997.

[12] J. Cole, H. Choi, H. Kim and M. Hasegawa-Johnson, "The effect of accent on the acoustic cues to stop voicing in Radio News speech," in *Proc. Internat. Conf. Phonetic Sciences*, 2003.

[13] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Current results," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1553-1573, April 1988.

[14] M. E. Beckman and J. Edwards, "Lengthenings and shortenings and the nature of prosodic constituency," in *Between the grammar and physics of speech: Papers in laboratory phonology I*, J. Kingston and M.E. Beckman (Eds), Cambridge: Cambridge University Press, pp. 152-178, 1990.

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

[16] M. E. Beckman and G. A. Elam, *Guidelines for ToBI labelling: the very experimental HTML version.* www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html, 1994.

[17] M. Ostendorf, B. Byrne, M. Fink, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley and T. Zeppenfield, "Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode," Report of the CSLU 1996 Summer Workshop.

[18] 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*.

[19] A. Dainora, "Eliminating downstep in prosodic labeling of American English," in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, pp. 41-46, 2001.

[20] S. Kim, M. Hasegawa-Johnson and K. Chen, "Automatic recognition of pitch movements using time-delay recursive neural network," in review.

[21] 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.