Prosody Dependent Speech Recognition on Radio News Corpus of American English

Ken Chen, Mark Hasegawa-Johnson, *Senior Member, IEEE*, Aaron Cohen, Sarah Borys, Sung-Suk Kim, Jennifer Cole, and Jeung-Yoon Choi

***Abstract*—Does prosody help word recognition? This paper proposes a novel probabilistic framework in which word and phoneme are dependent on prosody in a way that reduces word error rates (WER) relative to a prosody-independent recognizer with comparable parameter count. In the proposed prosody-dependent speech recognizer, word and phoneme models are conditioned on two important prosodic variables: the intonational phrase boundary and the pitch accent. An information-theoretic analysis is provided to show that prosody dependent acoustic and language modeling can increase the mutual information between the true word hypothesis and the acoustic observation by exciting the interaction between prosody dependent acoustic model and prosody dependent language model. Empirically, results indicate that the influence of these prosodic variables on allophonic models are mainly restricted to a small subset of distributions: the duration PDFs (modeled using an explicit duration hidden Markov model or EDHMM) and the acoustic-prosodic observation PDFs (normalized pitch frequency). Influence of prosody on cepstral features is limited to a subset of phonemes: for example, vowels may be influenced by both accent and phrase position, but phrase-initial and phrase-final consonants are independent of accent. Leveraging these results, effective prosody dependent allophonic models are built with minimal increase in parameter count. These prosody dependent speech recognizers are able to reduce word error rates by up to 11% relative to prosody independent recognizers with comparable parameter count, in experiments based on the prosodically-transcribed Boston Radio News corpus.**

***Index Terms*—Acoustic model, ANN, duration, HMM, mutual information, pitch, prosody, ToBI, word error rate.**

I. 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 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 can be described by a set of prosodic variables such as prosodic phrase boundary, pitch accent, lexical stress, syllable position and hesitation, etc. Among these prosodic variables, pitch accent and intonational phrase boundary have the most salient acoustic correlates, and may be most perceptually robust [2]. 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 [3]. 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 [4], while accented consonants are produced with greater closure duration [5], greater linguopalatal contact [6], longer voice onset time, and greater burst amplitude [7]. Knowledge of pitch accent placement would therefore be useful prior information for accurate acoustic modeling. Intonational phrase boundaries, which segment an utterance into intonational phrases, not only introduce a distinctive pitch contour (called a boundary tone) on the preceding speech segments, but also affect the acoustic realization of neighboring phonemes: phonemes preceding phrase boundaries are lengthened considerably [8] and consistently [9], phonemes both preceding and succeeding intonational phrase boundaries have more extreme lingual articulations [6], and vowels following an intonational phrase boundary are more likely to show a glottalized onset [10].

Prosody is potentially useful in automatic speech understanding systems for at least four reasons. First, prosody correlates with syntax: Price *et al.* [11] showed that prosody may be used to disambiguate syntactically distinct sentences with identical phoneme strings, while Kim *et al.* [12] have demonstrated that prosody may be used to infer punctuation of a recognized text. Second, prosody correlates with meaning: for example, Taylor *et al.* [13] 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 [14]. Finally, prosody may be useful as prior conditioning information for the correct phoneme labeling of an ambiguous acoustic signal.

This paper focuses on the fourth application of prosody: in this paper, prosody is used as prior conditioning information for the correct phoneme labeling of an ambiguous acoustic signal. We expect that with prosody accurately modeled in both acoustic model and language model, the word recognition performance will improve.

The correlation of prosodic and phonetic cues is well attested: for example, Cole *et al.* have shown that the voice onset time of a voiced stop in a pitch-accented syllable is comparable to that of an unvoiced stop in an unaccented syllable, thus the phoneme label is unambiguous only given prior information

Manuscript received July 1, 2003; revised August 29, 2004. This work was supported by the University of Illinois Critical Research Initiative and by NSF Award 0132900. Statements in this paper reflect the opinions and conclusions of the authors and are not endorsed by the NSF. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Bayya Yegnanarayana.

The authors are with Beckman Institute, University of Illinois, Urbana, IL 61801 USA (e-mail: kenchen@uiuc.edu).

Digital Object Identifier 10.1109/TSA.2005.853208

TABLE I
SUMMARY OF ALL THE PDF SPLITTING/CLUSTERING EXPERIMENTS CONDUCTED TO ASSESS THE PROSODIC EFFECTS ON THE PDFS OF THE ALLOPHONE MODELS

Experiments	Section	PDFs tested	Prosodic conditions
How boundaries affect phoneme duration	III-A	Duration PDFs	Intonational phrase boundary
How pitch accents affect acoustic-prosodic observations (f_0)	III-B	Acoustic-prosodic observation PDFs	Pitch accent
The joint effects of boundary and accent	IV-A	Duration PDFs and acoustic-prosodic observation PDFs	Intonational phrase boundary and pitch accent
How prosody affects the tree-based clustering of phonetic state observation PDFs (estimated from MFCCs)	III-C, IV-B	Acoustic-phonetic observation PDFs	Phonetic and prosodic conditions selected automatically by decision trees
Manual splitting of allophone models with complexity uncompensated (ULL), mixture-compensated (MLL) or triphone-compensated (TLL)	III-C, IV-B	Acoustic-phonetic observation PDFs	Phonetic and prosodic conditions selected manually based on their contribution to the average log likelihood scores

about the prosodic label [7]. On the other hand, a number of papers suggest that it may not be possible to uniquely determine the prosody of an utterance without prior knowledge of phoneme content: Wightman *et al.*, for example, suggest that prosodic phrase boundaries are best detected by comparing the expected and actual durations of phonemes in each word [15]. Since prosody is ambiguous without phoneme information, and phonemes are ambiguous without prosodic information, this paper describes a system that recognizes both at the same time. Specifically, we propose a set of prosody-dependent allophone models for speech recognition that effectively capture the influence of prosody at the phonetic level without significantly increasing the parameter count of recognizers. To measure the influence of prosody on the PDFs of allophone models, we conduct five prosody dependent allophone recognition experiments as listed in Table I. Details of these experiments and the results will be reported in the corresponding sections.

To train phonetic models that are aware of prosody, a large prosodically labeled speech database is required. However, hand labeling of prosody is known to be a difficult task even with a well formulated prosody labeling system [16]. Shriberg *et al.* [17]–[19] have proposed a different approach that makes use of acoustic prosodic cues without requiring explicit prosodic labeling. In their approach, phonological prosodic events (e.g., pitch accents and prosodic phrase boundaries) are not explicitly modeled. Instead, prosodic cues (e.g., pitch, energy) are conditioned over a set of hidden event variables representing sentence and topic boundaries, disfluency markers, dialog act labels etc. that are strongly correlated with prosodic phonological events. The advantage of their approach is that the hidden event labels are relatively easier to acquire than prosodic labels, making it possible to build large systems on standard speech corpora. The disadvantage of their approach is that the relationship between acoustic prosodic cues and syntax or disfluency is not as predictable as the relationship between acoustic cues and prosody [9]. As a result, their event dependent acoustic models can not directly utilize those well-attested prosodic phenomena: for example, pre-boundary lengthening [15]. Nonetheless, their prosody dependent systems have achieved better performance than prosody independent systems on a large scale spontaneous speech corpus.

The availability of the Boston University Radio News Corpus, one of the largest corpora designed for study of prosody [20] makes it possible for us to build linguistically meaningful prosodic models for speech recognition. 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 nonprofessional 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 prosodic labeling system represents prosodic phrasing, phrasal prominence and boundary tones, using the Tones and Break Indices (ToBI) system for American English [16]. 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. There are four primary types of intonational phrase boundary tones: L-L%, representing a declaration-final pitch fall, H-L%, representing a medial pitch in the middle of a longer declarative dialog turn, H-H%, representing a canonical yes-no question contour, and L-H%, representing a word-gap question; the contours !H-L% and !H-H% are less frequently observed. Seven types of accent tones are labeled: $\mathrm{H}^*, !\mathrm{H}^*, \mathrm{L}+\mathrm{H}^*, \mathrm{L}+!\mathrm{H}^*, \mathrm{L}^*, \mathrm{L}^*+\mathrm{H}$, and $\mathrm{H}+!\mathrm{H}^*$. 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 [20]. Presence versus absence of pitch accent is transcribed with 91% inter-transcriber agreement. Some accent label distinctions are more problematic than others: the L^* versus H^* distinction is quite robust, while the $\mathrm{L}+\mathrm{H}^*$ versus $\mathrm{L}^*+\mathrm{H}$ distinction is subject to considerable inter-transcriber disagreement.

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%). Dainora [21] argues that !H and H movements are not linguistically distinct and should therefore not be distinctly recognized. Taylor [22] argues further that L^* accents are prominent by virtue of their duration and possibly increased energy, but that they are not characterized by any pitch contour distinct from the connecting contour of nonaccented syllables; he advocates for a system in which all H^* and $!\mathrm{H}^*$ accents are classified as pitch events (e), and all L^* and unaccented syllables are nonevents (c). Taylor's system uses HMMs to model the pitch accents. When neural networks (including both time-delayed recurrent networks and feedforward networks) are used [23], we find that L^* pitch contour is recognized more often as H^* than as unaccented.

This paper is organized as follows: Section II presents the mathematic framework for a prosody dependent speech recognizer, consisting of word and allophone models dependent on two important prosodic variables: the intonational phrase boundary and the pitch accent. The goal of prosody-dependent word recognition is motivated by an information-theoretic analysis, resulting in an explicit statement of the conditions under which prosody-dependent allophone models and prosody-dependent language models jointly act to improve the posterior probability of the correct word string. In order to avoid large increases in the parameter count of the recognizer, this section proposes a parsimonious architecture in which all observation PDFs are tied into monophone or triphone classes, except for the explicit duration PDFs, the acoustic-prosodic observation stream (pitch), and a selected set of cepstral observations found to be most significantly affected by prosody. The strength of the prosody dependence in the allophone models are measured by various allophone recognition experiments. Section III considers the empirical evidence in support of specific prosody-dependent observation distributions. Prosody dependent recognizers that depend on only one prosodic variable (boundary or accent) are trained using explicit duration HMMs (EDHMMs) and HMMs to assess the individual influence of these prosody factors. The training and decoding algorithms of the EDHMM are given, together with the results of allophone recognition experiments. Based on the theoretical analysis in Section II and the empirical results in Section III, a prosody-dependent speech recognizer is trained and tested using the Radio News Corpus, and results are presented in Section IV. Conclusions are reviewed in Section V.

II. Prosody Dependent Speech Recognition

The task of speech recognition, given a sequence of observed acoustic feature vectors $O = (o_1, \ldots, o_T)$, is to find the sequence of word labels $W = (w_1, \ldots, w_M)$ that maximizes the joint probability $p(O, Q, W)$

$$[\tilde{W}] = \arg\max_W p(O, Q, W) \tag{1}$$

where $Q = (q_1, \ldots, q_L)$ is a sequence of sub-word units, typically allophones dependent on phonetic context. Ostendorf *et al.* [24] suggested expanding (1) as

$$[\tilde{W}] = \arg\max_W p(O|Q, H)p(Q, H|W, P)p(W, P) \tag{2}$$

where $P = (p_1, \ldots, p_M)$ is a sequence of prosody labels, one associated with each word, 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.

The models described in this paper may be understood as implementations of (2), with parameter count limited through selective implementation of prosody dependence using both acoustic-phonetic and empirical selection criteria. This section describes the information-theoretic motivation for prosody-dependent allophones, and the mathematical structures necessary to implement a parsimonious prosody-dependent acoustic model.

A. Information-Theoretic Analysis

Let a prosody-dependent allophone model be defined as an HMM whose states are conditioned on both phoneme label q_l and prosodic state h_l. Assume that a prosody-dependent pronunciation model may be pre-compiled so that each prosody-dependent word label $[w_m, p_m]$ corresponds to a unique hidden Markov model, created by concatenating an appropriate sequence of prosody-dependent allophone models. Let the prosody dependent language model be defined to be any standard language model (this paper will use bigram models) describing the probability of $[w_m, p_m]$ given the history $[w_1, p_1, \ldots, w_{m-1}, p_{m-1}]$.

The average modeled mutual information between the true word hypothesis W_T and the acoustic observation O may be defined as

$$I(O; W_T) = E_{W_T,O}\left\{\log \frac{p(O, W_T)}{p(O)p(W_T)}\right\} \tag{3}$$

where the expectation is computed over the true joint distribution of W_T and O, but the probabilities in the fraction are modeled probabilities; thus $I(O; W_T)$ is a measure of the quality of the PDF model $p(O, W_T)$. Suppose that $p(W_T)$ in (3) is defined to be the true probability of W_T, so that only the terms $p(O)$ and $p(O, W_T)$ depend on the quality of the speech recognition model. Under this definition, the quantity $I(O; W_T)$ is related by a constant to the model discriminant function $\Phi(O; W_T)$ [25], defined as

$$\begin{aligned}\Phi(O; W_T) &= E_{W_T,O}\{\log p(W_T|O)\} \\ &= E_{W_T,O}\left\{\log \frac{p(O, W_T)}{\sum_{\hat{W}} p(O, \hat{W})}\right\} \\ &= -E_{W_T,O}\left\{\log\left(\sum_i \eta_i\right)\right\}\end{aligned} \tag{4}$$

where

$$\eta_i = \frac{p(O, \hat{W}_i)}{p(O, W_T)} = \frac{p(O|\hat{W}_i)}{p(O|W_T)} \times \frac{p(\hat{W}_i)}{p(W_T)} \tag{5}$$

which is the likelihood ratio comparing the ith word sequence hypothesis $\hat{W}_i$ to the true word sequence W_T.

The discriminant function of a prosody dependent recognizer can be represented as

$$\Phi_P(O; W_T) = -E_{W_T,O}\left\{\log\left(\sum_i \acute{\eta}_i\right)\right\} \tag{6}$$

where

$$\begin{aligned}\acute{\eta}_i &\approx \frac{\max_{\hat{P}} p(O, \hat{W}_i, \hat{P})}{\max_{\hat{P}} p(O, W_T, \hat{P})}, \\ &= \frac{p(O|\hat{W}_i, \hat{P}_i)}{p(O|W_T, P_T)} \times \frac{p(\hat{W}_i, \hat{P}_i)}{p(W_T, P_T)}\end{aligned} \tag{7}$$

where P_T is the prosody sequence that maximizes $p(O, W_T, \hat{P})$, and $\hat{P}_i$ is the prosody hypothesis that maximizes $p(O, \hat{W}_i, \hat{P})$.

The objective of prosody-dependent speech recognition in this paper is to create prosody-dependent speech recognition models such that $\Phi_P(O;W_T) > \Phi(O;W_T)$, thus increasing the modeled probability of the correct word sequence given the observation. From (4) and (6), $\Phi_P(O;W_T) > \Phi(O;W_T)$ if

$$E_{W_T,O}\left\{\log\left(\frac{\sum_i \acute{\eta}_i}{\sum_i \eta_i}\right)\right\} < 0. \tag{8}$$

Equation (8) expresses the condition under which prosody-dependent speech recognition increases the modeled mutual information $I(O;W_T)$. In order to guide the design and interpretation of experiments in the field of prosody-dependent speech recognition, it is valuable to spend some time trying to express the meaning of (8) in words. Loosely speaking, (8) claims that modeled mutual information improves if $\acute{\eta}_i < \eta_i$ for most combinations of W_T and $\hat{W}_i$, where the word "most" is quantified by the expectation over W_T of the log ratio of sums over $\hat{W}_i$. Re-arranging terms, the condition $\acute{\eta}_i < \eta_i$ may be written

$$\left(\frac{p(P_T|W_T)}{p(\hat{P}_i|\hat{W}_i)}\right)\left(\frac{p(O,W_T|P_T)/p(O,W_T)}{p(O,\hat{W}_i|\hat{P}_i)/p(O,\hat{W}_i)}\right) > 1. \tag{9}$$

Equation (9) expresses the fraction $\eta_i/\acute{\eta}_i$ as the product of two terms.

The first term on the left expresses the improvement, due to prosody, in the selectivity of the language model. It is positive, for example, when the true word sequence is uttered with a highly predictable prosodic pattern, thus $p(P_T|W_T) > p(\hat{P}_i|\hat{W}_i)$. This term may be maximized by modeling only those prosodic labels that are most predictable from word sequence statistics. In this paper, prosodic labeling will include intonational phrase boundaries and phrasal pitch accent. Previous research [26], [27] has shown that intonational phrase boundaries are well predicted by N-gram word sequence statistics.

The second term on the left expresses the improvement, due to prosody, in the selectivity of the acoustic model. It is positive, for example, when the observation sequence O is better explained by P_T than by $\hat{P}_i$. This term may be maximized by selectively modeling only those acoustic features whose distributions are well predicted by prosodic labeling. Beckman *et al.* [3] suggest that talker-normalized fundamental frequency (f_0) is well predicted by the location of pitch accents, while Wightman *et al.* [15] suggest that normalized phoneme duration is well predicted by the location of intonational phrase boundaries. Cole *et al.* [7] describe the prosody-dependent modification of the acoustic-phonetic features (e.g., MFCC) as a reliable effect in the case of some phonemes but not all phonemes, thus prosody-dependent modification of the distribution of MFCCs will be modeled only for an empirically selected subset of phonemes.

The meaning of (8) may therefore be explained in the following words: $\Phi_P(O;W_T) > \Phi(O;W_T)$ if, most of the time, the correct prosodic sequence is well predicted by the word transcription, and the acoustic observation is well predicted by the prosody. Note that it is possible for a prosody-dependent speech recognizer to result in reduced word error rate 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.

B. Prosody Dependent Allophone Models

Equation (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 (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 (2) is rarely justified by a proportional increase in recognition accuracy. To increase the trainability of the models and reduce the computational cost, we propose to model only a subset of phonetic distributions that are known to be most sensitive to prosodic context. Specifically, we propose to model the prosody dependence of the phonetic state duration PDFs, which is known to be affected significantly by intonational phrase boundary, and the acoustic-prosodic observation PDF, which models the distribution of the acoustic observation of pitch accents. The acoustic-phonetic observation PDFs, that is, the PDFs that model the spectral distribution of the phonetic states, ignore prosody by default; only a small set of acoustic-phonetic PDFs are allowed to depend on prosody. By limiting the effect of prosody in this way, we create effective models of the most striking and most often reported prosody-dependent allophonic variation, without significantly increasing the parameter count of the speech recognizer.

In the proposed system, the prosody state variable h can be represented as a two dimensional prosodic vector: $h = [a, b]$, where a is a discrete variable indicating the pitch accent prominence level of q, and b is another discrete variable indicating the lengthening level of q as affected by the intonational phrase boundaries. The observation vector O usually contains only acoustic phonetic observation X (typically cepstral coefficients) that provide cues for the discrimination of phonetic units. In the prosody dependent framework, it can be augmented to include an additional acoustic-prosodic observation Y that contains features (typically pitch) as acoustic cues for the detection of prosodic events: $O = [X, Y]$.

In this paper, phrase-final lengthening is modeled by conditioning the state residency time or "duration" of an HMM state on the prosodic variable b. The pitch accent is modeled by conditioning the distribution of the acoustic-prosodic observation Y on both the phoneme state q and the prosodic variable a. In order to precisely model prosody dependent phoneme lengthening, we propose to use a prosody-dependent explicit duration hidden Markov model (EDHMM). The EDHMM of phoneme q_i under prosodic state b_i consists of a sequence of hidden phonetic state variables $S_i = (s_{i1}, \ldots, s_{iN})$, 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 the distribution of duration and acoustic prosodic observation, then the probability of observing matrix $O_i = [O_{i1}, \ldots, O_{iN}]$ is

$$\begin{aligned} p(O_i|q_i,h_i) &= p(X_i,Y_i|S_i,h_i)p(S_i|q_i,h_i) \\ &= \prod_{j=1}^{N} p(X_{ij},Y_{ij}|s_{ij},a_i)p(d_{ij}|s_{ij},b_i)p(S_i|q_i) \\ &= \prod_{j=1}^{N} p(X_{ij}|s_{ij})p(Y_{ij}|s_{ij},a_i)p(d_{ij}|s_{ij},b_i)p(S_i|q_i). \end{aligned} \tag{10}$$

Equation (10) proposes a parameter sharing strategy across different prosody dependent allophonic models $[q_i, h_i]$: the state acoustic-phonetic observation PDFs $p(X_{ij}|s_{ij})$ are shared regardless of their prosodic status a or b, the state duration PDFs are shared if they are conditioned on the same value of b (the same level of lengthening), and the state acoustic-prosodic observation PDFs are shared if they are conditioned on the same value of a (the same level of pitch accent prominence). This way of parameter sharing reflects the phonetic knowledge of how prosody affects the duration, pitch and spectral distribution. It helps us constrict the complexity of the acoustic model and avoid significant increases in parameter count. It is possible to model multiple levels of prosodic boundaries and prominences using this framework (i.e., increasing the cardinalities of a and b). However, modeling more prosodic distinctions inevitably fragments the training data and makes our approach impractical on small corpora. Therefore, we decide that a and b only take binary values in our final system, which in combination, create four prosody-dependent allophones of each phonetic model q_i, namely: neutral (default), accented, lengthened, and $\mathrm{accented} + \mathrm{lengthened}$.

C. Explicit Duration HMM

In order to enable prosody-dependent modeling of duration, the hidden Markov toolkit (HTK) was modified in order to implement a variant of Ferguson's explicit duration hidden Markov model (EDHMM) [28]. In the literature, there are two successful algorithms that explicitly model HMM duration as random variables by extending the underlying Markov chain to a semi-Markov chain. Ferguson [28] first proposed an Estimation Maximization (EM) algorithm to estimate a nonparametric probability mass function (PMF) for the duration of each state. Levinson [29] later proposed the continuously variable duration HMM (CVDHMM) in which the state duration probability is modeled as a continuous gamma density function. As compared with Levinson's algorithm, Ferguson's algorithm requires a large amount of training data but has no prior assumption on the parametric form of the duration distribution. Ferguson's algorithm is also more computationally efficient than Levinson's algorithm: Ferguson's algorithm requires $O(NT(N+D))$ operations during training, in contrast to $O(N^2TD^2)$ operations required by Levinson's algorithm, where N is the number of states in the HMM, T is the total number of observations in the example, and D is the maximum allowed state duration.

The details of the algorithms for explicit duration HMMs, required in order to perform both training and recognition search, are described in the Appendix. The efficiency of the training algorithm makes it practical to train EDHMMs on a large speech corpus in an amount of time comparable to standard HMMs. To verify the performance of EDHMM for prosody-independent phoneme modeling, phoneme recognition experiments were conducted on the TIMIT database. 48 phonemes were each modeled by a 3-state HMM with 3 mixture Gaussians per state, and with no language model. Observations included energy, fifteen MFCCs, and their delta coefficients once per 10ms. Phoneme recognition experiments using a standard HMM without explicit duration models resulted in 51.0% accuracy; phoneme recognition experiments using the EDHMM resulted in 51.9% accuracy. The explicit duration model increases the total parameter count of each HMM state by D trainable parameters. The maximum allowed state duration D is chosen automatically by restricting the dynamic range of the duration PMF, but is typically $5 \leq D \leq 15$. In the experiments reported here, the EDHMM requires roughly 5% more trainable parameters than the HMM. In our experiments, we did not find that a 5% increase in the parameter count required any corresponding increase in the size of the training database.

D. Prosody-Dependent Word Transcription

In experiments described in this article, the language model $p(W,P)$ is implemented as a prosody-dependent bigram, i.e.,

$$p(W,P) = p(w_1,p_1)\prod_{m=2}^{M} p(w_m,p_m|w_{m-1},p_{m-1}). \tag{11}$$

The prosodic label p_m carries two types of information: the pitch accent status of word w_m, and the position of w_m within an intonational phrase. There are eight possible settings of p_m: a word may be accented or unaccented; the same word may be phrase-initial, phrase-final, phrase-medial, or it may be a one-word intonational phrase (both phrase-initial and phrase-final). Equation (11) is implemented by defining a unique symbol corresponding to each possible label vector $[w_m, p_m]$. A prosody-dependent word transcription may contain prosody-dependent word tokens of the form W_ab, where W is the word label, a takes the values "a" or "u" (accented or unaccented), and b takes the values "i, m, f, o" (initial, medial, final, one-word phrase). In this scheme, 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."

The sequence $[p_{m-1}, p_m]$ takes on $|P|^2 = 64$ possible values, so in theory, a prosody-dependent bigram model learns 64 times as many parameters as a prosody-independent bigram model. In practice, most possible combinations of w_m and p_m never occur, so their probabilities are estimated by backing off to 1-gram and 0-gram (uniform) distributions; in our experiments, the actual parameter count of a prosody-dependent bigram model is slightly less than three times that of a prosody-independent bigram.

The pronunciation model $p(Q,H|W,P)$ is implemented using a prosody-dependent dictionary. A prosody-dependent dictionary is a lookup table providing the prosody-dependent allophone pronunciation of each prosody-dependent word. Prosody-dependent allophones are tagged in the same way as

prosody-dependent words, i.e., a fully-specified prosody-dependent allophone may have the form P_ab, where P is the monophone or triphone label, specified in SPHINX notation [30]. Experiments reported in Section III test the importance of the prosodic variables both together and in isolation; different prosody-dependent dictionaries were designed for each experiment, with entries matched to the prosodic variables under test. For example, to model phrase-final lengthening effects while ignoring all other prosodic effects, the final vowel (FV) and final coda consonants (FC) in a phrase final word (W_af, W_uf, W_ao, or W_uo) are labeled as phrase-final (P_f), while other phones are labeled as phrase-medial (P_m).

A phrase-level pitch accent prominence on a multisyllabic word usually falls on or near the syllable with primary lexical stress. Exceptions to this rule include emphatic accent, which may lengthen the entire word, and contrastive accent, which may be applied to a syllable other than the primary stress syllable. In order to limit the complexity of the recognition model, experiments reported in this paper ignore these special cases; instead, the dictionary entry of an accented word contains accented allophone models only in the syllable with primary lexical stress. Thus, for example, a fully-specified dictionary entry for "wanted_af" contains the allophone list "w_am aa_am n_am t_um ix_uf d_uf."

III. Modeling the Prosody Induced Allophonic Variation

A large number of experiments were conducted in order to specify, as precisely as possible, the effect of prosody on the observation likelihoods of a speech recognition model. This section describes results of these experiments. In particular, this section will focus on three possible effects of prosody: the effect of phrase position on allophone duration, the effect of pitch accent on the acoustic-prosodic observations (pitch), and the effect of any prosodic variable on the acoustic-phonetic (cepstral) observations. The goal of these experiments is to determine exactly which observation probability densities should be modeled as prosody-dependent, given the restriction that total parameter count of the model must not be substantially increased.

A. Phrase-Boundary Lengthening

Prosody-dependent allophone recognition experiments were conducted on the Radio News Corpus using allophone models dependent on intonational phrase position (b) but not pitch accent (a). In this case, (10) reduces to

$$p(O_i|q_i,h_i)=\prod_{j=1}^{N}p(X_{ij}|s_{ij})p(d_{ij}|s_{ij},b_i)p(S_i|q_i). \quad (12)$$

Eight experiments compare each of the eight IPB-dependent allophone sets in Table II to a matched boundary-independent (BI) allophone set. In experiments using the "final vowel" and/or "final consonant" allophone sets (FV, FC, FVFC), vowels and/or coda consonants in the syllable before an IPB are marked $b=\text{phrase-final}$; all other allophones are phrase-medial. In experiments using the IV, IC, or IVIC allophone sets, vowels and/or onset consonants in the syllable following an IPB are marked $b=\text{phrase-initial}$, and others are phrase-medial. The ICFV and IPFP sets include phrase-initial, phrase-medial, and phrase-final allophones.

TABLE II
Eight Different Intonational-Phrase-Boundary Context Definitions Were Tested, With Different Subsets of Phonemes Allowed to Be Either Phrase-Initial or Phrase-Final. Next to Each Context Definition Are Listed the Resulting Number of Allophones, the Number of Parameters in an HMM-Based Recognizer, and the Number of Parameters in an EDHMM Recognizer

	Lengthened allophones	**# Phn**	**HMM # Params**	**EDHMM # Params**
BI	None	65	39065	42093
FV	phrase Final Vowels	89	39170	42713
FC	phrase Final Consonants	91	39240	42824
FVFC	FV+FC	105	39345	43519
IV	phrase Initial Vowels	87	39247	42713
IC	phrase Initial Consonants	83	39219	42784
ICIV	IC+IV	102	39401	43462
ICFV	IC+FV	98	39303	43380
IPFP	ICIV+FVFC	153	39688	44718

TABLE III
IPB Dependent Allophone Recognition Accuracy (%) With Each Allophone Counted as a Distinct Token

	HMM		**EDHMM**	
	BI	BD	BI	BD
FV	25.70	33.93	26.10	34.36
FC	13.22	27.4	13.61	28.02
FVFC	3.13	24.61	3.77	25.36
IC	28.55	25.53	29.28	25.92
IV	31.95	30.09	32.45	30.77
IVIC	23.15	19.10	23.57	19.71
ICFV	23.88	22.89	24.28	23.20
IPFP	1.71	12.19	2.35	12.91

For each of the prosodic context definitions in Table II, two sets of allophone models were constructed: an IPB-dependent set BD, and a baseline IPB-independent set BI. Both BD and BI contain the same set of allophones. In BI, the allophones created from any given monophone share all parameters and are in fact identical, whereas in BD, allophones based on the same monophone share only observation PDFs but have independent duration PDFs. By comparing the allophone recognition accuracy of BD and BI models with a null grammar (every allophone sequence equally likely), it is possible to assess the strength of the dependence between phoneme duration and each type of prosodic context defined in Table II. Table III shows results of this experiment. Note that figures in different rows are not comparable because they are measured under different prosody contexts with allophone sets of different sizes.

Table III shows that distinctive modeling of phrase-final phoneme duration PDFs (FV, FC, and FVFC conditions) significantly improves allophone recognition accuracy for both HMMs and EDHMMs (note that in this result, allophones $[q,h]$ based on the same monophone q are counted as different symbol if h is different). Wightman *et al.* [15] found that phonemes in intonational phrase final rhymes are significantly longer than

similar phonemes in other contexts; Table III indicates that the transition probabilities of an HMM and the explicit duration PMF of an EDHMM are both capable of learning the distinction between phrase-final and nonphrase-final phonemes. Conversely, distinctive modeling of phrase-initial phoneme duration PDFs (IV, IC, and IVIC conditions) degrades allophone recognition accuracy, indicating that HMM and EDHMM fail to find any systematic duration variation at phrase-initial position. Non-phrase-initial phones usually appear much more often than do their phrase-initial counterparts. Therefore, unless position dependent duration modeling increase the accuracy of both phrase-initial models and nonphrase-initial models, worse phoneme recognition performance is expected. It can be concluded from this experiment that IPB dependent lengthening only happens at phrase-final positions including both final vowels and final consonants.

B. Accent Dependent Phonetic Modeling

Allophonic variation induced by pitch accent was investigated using HMMs and EDHMMs trained to recognize a nonlinearly transformed acoustic-prosodic observation vector.

Fundamental frequency f_0 is extracted from speech by using the `formant` program in Entropic XWAVES. The probability of voicing (PV) is output at the same time as a confidence measure to the extracted f_0. Measured f_0 values are then smoothed and normalized using the following algorithm. Like most automatic pitch trackers, this raw f_0 usually contains some amount of pitch doubling and halving errors. To avoid these pitch doubling and halving errors, a 3 mixture Gaussian classifier, based on the method proposed by Sonmez [31] is trained on the f_0 data from each utterance, with mixture component means constrained to equal $1/2$, 1, and 2 times the utterance mean pitch, and measured f_0 candidates classified as apparently equal to $2f_0$ or $f_0/2$ are eliminated. f_0 measurements with small PVs are also eliminated, because frames with small PV are usually unvoiced and result in unreliable pitch measurements. Remaining f_0 measurements are normalized and converted to log scale using the formula

$$\hat{f}_0 = \log\left(\frac{f_0}{\mu} + 1\right) \tag{13}$$

where μ is the utterance mean pitch. Frames with missing $\hat{f}_0$ are filled by linearly interpolating $\hat{f}_0$ between available frames, resulting in a smoothed normalized pitch waveform $\tilde{f}_0$.

$\tilde{f}_0$ does not necessarily have a Gaussian distribution. To use it as an acoustic feature for HMM, we need to transform it to a new variable Y with the following characteristics: (1) the values of Y observed during accented and unaccented syllables are as distinct as possible, and (2) the distribution of Y is reasonably well modeled by a Gaussian PDF (in our study, we find that the class-conditional histograms of Y over the training data have a bell shape). These goals are well approximated by a multi-layer perceptron (MLP) trained to accept five consecutive measurements of $\tilde{f}_0$ as input, and to estimate a nonlinear function $Y(t) = g([\tilde{f}_0(t-2), \ldots, \tilde{f}_0(t+2)])$ such that the values of $Y(t)$ resulting from accented and unaccented frames are maximally distinct. Comparing the signals before and after the transformation, we find that this type of nonlinear transformation can help reduce the pitch declination effect [32]. In addition, it is known that pitch accents are often not aligned with the perceptually prominent syllable. This type of minor pitch-peak asynchrony can also be modeled by the MLP, which computes the acoustic-prosodic observation based on a summary of five consecutive f_0 observations.

TABLE IV
PERCENT CORRECTNESS AND ACCURACY FOR ACCENT DEPENDENT ALLOPHONE RECOGNITION WITH EACH ALLOPHONE COUNTED AS A DISTINCT SYMBOL. COLUMN 2 IS THE TYPE OF THE ACOUSTIC PROSODIC FEATURE USED IN TRAINING AND TESTING AND COLUMN 3 IS THE NUMBER OF MIXTURES USED FOR THE ACOUSTIC PROSODIC OBSERVATION PDFS

	A. P. Feature	# Mix	Corr(%)	Acc(%)
AI	None	0	30.81	16.13
AD1	Y	3	37.21	21.93
AD2	Y	1	37.16	21.86
AD3	$\tilde{f}_0$	3	35.66	20.34

If the setting of the phrase-boundary variable b is ignored, (10) reduces to

$$p(O_i|q_i,h_i) = \prod_{j=1}^{N} p(X_{ij}|s_{ij}) \times p(Y_{ij}|s_{ij},a_i)p(d_{ij}|s_{ij})p(S_i|q_i). \tag{14}$$

HMM recognizers were trained for the purpose of accent-dependent allophone recognition. Three recognizers were designed using the same set of labels: the label set includes one accented and one unaccented version of each monophone in the SPHINX phoneme set. In the accent-independent (AI) recognizer, allophones of the same monophone are physically identical: all model parameters are shared. Each allophone pair in AD1 and AD2 has tied acoustic-phonetic (cepstral) observation PDF but untied acoustic-prosodic ($Y = g(\tilde{f}_0)$) observation PDF. AD1 and AD2 differ in that AD1 uses 3 mixture Gaussians for the acoustic prosodic observation PDF, whereas AD2 only uses a single Gaussian. AD3 is the same as AD1 except that it is trained using $\tilde{f}_0$ as acoustic-prosodic features without nonlinear transformation. Allophone recognition results with no grammar are listed in Table IV. AD1 and AD2 yield similar recognition results in this experiment, indicating that the acoustic-prosodic observation PDFs can be well approximated by a single Gaussian. AD3 yields worse allophone recognition accuracy (20.34%) than does AD1, indicating the effectiveness of the MLP-based nonlinear transformation.

C. Influence of Prosody on MFCCs

The influence of prosody on the distribution of acoustic phonetic observations (typically MFCC), if it can be modeled, would further help increase the distinction among the prosody dependent allophones. It is worthwhile to determine whether or not such prosody dependent spectral variation can be modeled in this corpus to improve the accuracy of prosody dependent allophones, as phonetic study [4]–[7] has suggested that the distribution of spectral energy can be greatly affected by prosodic context.

This aspect of prosody research has been suggested by Lee in [30] after he successfully modeled the phonetic context dependent spectral variation using triphone models. Lee pointed out

that when more training data are available, triphone systems that cluster phonetic models based only on phonetic context should be extended to include additional sources of phonetic variability, such as syllable position, stress, nonneighboring phones, or interword triphones. Shafran *et al.* [33] applied a decision-tree algorithm on a subset of the Switchboard corpus to assess the influence of various prosodic factors on the phonetic state observation PDFs of a set of triphones. They found acoustic differences primarily associated with segment position at prosodic constituent onsets and in prominent syllables, and suggested that prosody-dependent phonetic models should be developed once a sufficient amount of prosodically-labeled data is available.

In our investigation, we conducted two different types of experiments on the Radio New Corpus to answer the question: are there prosody dependent spectral variations that can be modeled to improve the accuracy of allophone models in this corpus?

The first experiment evaluated prosodic questions in a tree-based allophone clustering algorithm, similar to the algorithms used by Shafran *et al.* [33]. Two sets of trees were constructed: a prosody-dependent tree (TPD) with prosodic questions available to the allophone clustering algorithm, and a prosody-independent tree (TPI) that used only phoneme context questions in order to perform allophone clustering. Each tree clusters the state-dependent observation PDFs of the SPHINX monophones according to an informatic-theoretic measure [34]. TPI is the standard triphone state clustering tree which asks only questions regarding the phonetic context of the given monophone (for example, the place or manner of articulation of the neighboring phones); whereas TPD asks questions regarding both phonetic context and prosodic context. The two experiments were controlled so that both the TPI and TPD systems result in the same total number of leaf nodes, that is, the recognizer is trained to learn the same total number of acoustic-phonetic observation PDFs, regardless of whether or not prosody is used in the state-tying tree. Experiments were conducted using 750, 1000, and 1250 total leaf nodes (these numbers are selected arbitrarily in order to measure the influence of prosodic questions under different conditions). If some questions regarding prosodic context have a larger influence on the state spectral distribution than do the questions regarding phonetic context, they are going to appear closer to the root of the tree, and the monophone recognition accuracy based on TPD should be greater than the monophone recognition accuracy based on TPI.

Only about half of the Radio News Corpus is prosodically transcribed; for the remainder of the corpus, transcriptions specify each word's part of speech and lexically stressed syllable, but not the locations of pitch accents and intonational phrase boundaries. The tree-based clustering algorithm requires a large amount of data, therefore it was trained using the entire Radio News Corpus. The clustering algorithm was therefore not able to distinguish allophones on the basis of pitch accent or intonational phrase position. The only prosodic questions available to the algorithm were 1) is the center/left/right phoneme lexically stressed or unstressed and 2) is the center/left/right phoneme part of a content word or a function word. Strictly speaking, the second type of questions are not prosodic but rather syntactic. However, they have strong correlation with prosodic phenomena as function words are unlikely to be accented while content words are likely to be accented.

Phoneme recognition accuracies of the TPD and TPI models will be reported below in Section IV. It will turn out that prosody dependence aids phoneme recognition accuracy; this result is not surprising, since the set of questions available to the TPD clustering algorithm is a strict superset of the set of questions available to TPI (i.e., all questions asked in TPI clustering are asked in TPD clustering but not vice versa). A closer analysis of TPD, however, shows that questions regarding prosody are used in only about 24% of the trees. The percentage of trees using lexical stress distinctions increases with increasing parameter count, from 17.1% in a system with 750 leaf nodes to 21.4% in a system with 1250 leaf nodes; the percentage of trees using the function/content part of speech distinction is relatively independent of parameter count, with 11.2–12.4% of the trees in the 750, 1000, and 1250-leaf systems. The percentage of trees using either one or both types of distinction grows with increased parameter count, from 23.6% of trees in the 750-leaf system to 28.1% of trees in the 1250-leaf system. Apparently the use of information about lexical stress and part of speech aids speech recognition performance, but the advantage seems to be significant only for about 24% of phonemes; for most phonemes, the clustering algorithm chooses phoneme context questions in preference to prosodic questions. Examples of phonemes affected by prosody include back vowels and nasal consonants: back vowels are affected by their own lexical stress (possibly resulting from the vowel reduction effect), while nasal consonants are affected by the lexical stress of neighboring vowels (possibly resulting from vowel nasalization and nasal flapping). This result is partially consistent with Shafran *et al.* [33], in which they report that among all the consonants, stop /t/ and nasal /n/ have the greatest variety given different prosody.

The high data requirements of the first experiment inspired a second set of experiments with similar objectives, but with parameter counts very tightly constrained. Consider a monophone q that may be split into two allophones q_1 and q_2 according to any criterion, including either a prosodic criterion (e.g., accent) or a phonetic criterion (e.g., manner of articulation of a neighboring phoneme). Let Λ_q represent an HMM trained to represent the monophone, while Λ_1 and Λ_2 are trained to represent the two allophones. Let D_q be the number of parameters in model Λ_q, while D_1 and D_2 are the parameter counts of models Λ_1 and Λ_2. A measure of the quality of the model Λ_q may be obtained by computing the average log probability of N_q tokens in a test database independent of the training set, thus

$$l(O_q|\Lambda_q) = \frac{1}{N_q} \sum_{i \in O_q} \log p(O_i|\Lambda_q) \tag{15}$$

where O_q is the set of test tokens, of size N_q. The measure $l(O_q|\Lambda_q)$ may be compared to the average log probability of the same test data, given the binary distinction Λ_1 versus Λ_2:

$$l(O_q|\Lambda_1, \Lambda_2) = \frac{1}{N_1} \sum_{i \in O_1} \log p(O_i|\Lambda_1) + \frac{1}{N_2} \sum_{i \in O_2} \log p(O_i|\Lambda_2) \tag{16}$$

where it is assumed that O_q is the union of O_1 and O_2, and $N_q = N_1 + N_2$. Equations (15) and (16) represent comparable log likelihood measures of two different recognition models: one model is given by the parameter set Λ_q, while the other

model is given by the union $[\Lambda_1, \Lambda_2]$. Thus the following equation may be used to test the justifiability of an allophone split. The allophone split $q \rightarrow [q_1, q_2]$ is said to be "justified" by a particular experiment if

$$l(O_q|\Lambda_1, \Lambda_2) > l(O_q|\Lambda_q) + B(N_q, D_1, D_2, D_q) \qquad (17)$$

where $B(N_q, D_1, D_2, D_q)$ is a threshold for significant improvement given by the Bayesian Information Criterion (BIC) [35]

$$B(N_q, D_1, D_2, D_q) = (D_1 + D_2 - D_q)\frac{\log N_q}{N_q}. \qquad (18)$$

Notice that (17) differs from most applications of the BIC in that it is computed using development test data, rather than being computed using the model training data.

Three experiments were conducted using model comparisons based on (15) and (16): an uncompensated log-likelihood experiment (ULL), a mixture-compensated experiment (MLL), and a triphone-compensated experiment (TLL). In the uncompensated experiment, the parameter counts were set to $D_1 = D_2 = D_q = 495$ parameters per model (3 states $\times$ 3 mixture components per state $\times$ 55 parameters per mixture), so that the BIC threshold was set to $495 \log N/N$ for a model with N training tokens. The MLL and TLL experiments use different techniques to set $D_1 + D_2 = D_q$, so that the BIC threshold is zero. The MLL experiment sets $D_1 + D_2 = D_q$ by giving model Λ_q twice as many Gaussian mixture components as Λ_1 or Λ_2. The triphone-compensated experiment (TLL) splits Λ_q into either prosody-independent allophones $[\Lambda_{\mathrm{PI1}}, \Lambda_{\mathrm{PI2}}, \Lambda_{\mathrm{PI3}}]$ or prosody-dependent allophones $[\Lambda_{\mathrm{PD1}}, \Lambda_{\mathrm{PD2}}, \Lambda_{\mathrm{PD3}}]$, where the models Λ_{PI} are defined by phoneme context, and the models Λ_{PD} may be defined by either phoneme or prosodic context. The total parameter counts are equal: $N_{\mathrm{PI1}} + N_{\mathrm{PI2}} + N_{\mathrm{PI3}} = N_{\mathrm{PD1}} + N_{\mathrm{PD2}} + N_{\mathrm{PD3}}$.

The uncompensated experiment (ULL) tested two prosodic questions: accent, and intonational phrase position [36]. Each monophone in the SPHINX monophone set was split into either two allophones ($\Lambda_1 = \mathrm{accented}, \Lambda_2 = \mathrm{unaccented}$) or three allophones ($\Lambda_1 = \mathrm{phrase\text{-}initial}, \Lambda_2 = \mathrm{phrase\text{-}medial}, \Lambda_3 = \mathrm{phrase\text{-}final}$). Equation (17) was satisfied for almost all possible vowel allophones, thus it was impossible to rule out any prosodic allophone of any vowel using this experiment. Equation (17) was not satisfied for all consonant allophones. Analysis of the results showed that the majority of consonants show acoustic evidence for at most three distinct prosodic allophones. First, the distinction between phrase-final accented and phrase-final unaccented consonants is clearly unjustified ((17) is not satisfied). Second, the distinction among phrase-initial accented, phrase-initial unaccented, and phrase-medial accented consonants is clearly unjustified. Third, both of these categories seemed to be distinct from the category of unaccented phrase-medial consonants. These three categories correspond reasonably well with categories that have been labeled in articulatory studies [6] as "lengthened (phrase-final)," "strengthened (accented or phrase initial)," and "default" consonant articulations, respectively.

The mixture-compensated experiment (MLL) evaluated four prosodic variables: pitch accent (accented versus unaccented), lexical stress (stressed versus unstressed), intonational phrase position (initial versus medial versus final), and part of speech (function word versus content word). The number of Gaussian mixtures in Λ_q was increased so that $D_q = D_1 + D_2$; otherwise this experiment was identical to the ULL experiment. The results of ULL and MLL were quite different. Equation (17) was satisfied for almost none of the allophone pairs considered in the MLL experiment. The only exception (the only prosodic allophone distinction that is certainly justified by the MLL experiment) is the distinction between accented and unaccented function-word vowels (such as /ax/ in "the"). This is a quite intuitive result, meaning that accent condition significantly affects the acoustic realization of function words.

The triphone-compensated experiment (TLL) evaluated three prosodic variables (accent, intonational phrase position, and part of speech), in addition to a list of variables encoding various binary distinctions among manner categories taken by the left and right context phones. In this experiment, the prosody-independent recognizer contained exactly three allophones of each SPHINX monophone, $[\Lambda_{\mathrm{PI1}}, \Lambda_{\mathrm{PI2}}, \Lambda_{\mathrm{PI3}}]$, created by splitting the monophone according to the two phoneme context variables with the highest contributions to log likelihood. The prosody-dependent recognizer also contained exactly three allophones of each SPHINX monophone, $[\Lambda_{\mathrm{PD1}}, \Lambda_{\mathrm{PD2}}, \Lambda_{\mathrm{PD3}}]$, created by splitting the monophone according to the two phoneme context or prosodic variables with the highest contributions to log likelihood. The model comparison is therefore compensated ($N_{\mathrm{PI1}} + N_{\mathrm{PI2}} + N_{\mathrm{PI3}} = N_{\mathrm{PD1}} + N_{\mathrm{PD2}} + N_{\mathrm{PD3}}$), but compensation is performed in such a way that the set of distinctions allowed by the prosody-dependent recognizer is a strict superset of the set of distinctions allowed by the prosody-independent recognizer.

The TLL experiment is similar in many ways to the tree-based HMM state clustering experiment. The most important difference between the two experiments is that, because the TLL experiment could be performed using a smaller database than the TPD experiment, it was possible to use a variety of prosodic questions that required accurate ToBI transcriptions of the data, including questions about pitch accent and intonational phrase position. As a result of this increased prosodic flexibility, a larger percentage of the resulting allophone definitions are prosody-dependent. In the TLL experiment, for eleven vowels and diphthongs (/ae,ah,aw,ay,ax,eh,ey,er,ih,iy,ow/), the most important context questions was pitch accent, and the second most important question was intonational phrase position. Among the vowels with enough tokens for evaluation, the only vowels that were more sensitive to phone context than prosodic context were /uw/ and /aa/. Both /uw/ and /aa/ were more sensitive to the manner of articulation of the right context phone than to any prosodic distinction; specifically, both phones were sensitive to the three-way distinction between stop, liquid, and other right context phones. Among consonants, only the unvoiced affricate and stops (/ch,p,t,k/) were more sensitive to prosodic than phonetic context variables. All four were sensitive to the phrase-final versus nonfinal distinction, and this distinction was most important for the phones (/p,t,k/), perhaps because the aspiration segments of nonphrase-final stops tend to be heavily co-articulated glide-like segments, while those of final stops are often acoustically distinctive noise bursts. All four phonemes were also sensitive to the distinction between accented and unaccented syllables (a prosodic distinction) and to the distinction between vocalic and nonvocalic right context

phones (a triphone context distinction); /k/ was more sensitive to the latter of these two questions, while /p/,/t/, and /ch/ were more sensitive to the former.

IV. Experiments and Results

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

A. Modeling of Duration and Pitch

Recognition experiments were conducted in order to test a recognizer with no prosodic distinctions among the acoustic-phonetic observation PDFs. In these experiments, the acoustic model $p(X, Y|Q, H)$ contained only two types of information about prosody. First, HMM transition probabilities and EDHMM duration PMFs were allowed to depend on the phrase position of an allophone. Second, the acoustic-prosodic PDF $p(Y|Q, H)$ explicitly modeled the distribution of nonlinearly transformed pitch features given phoneme label and pitch accent status.

In all experiments, a 3-state HMM with no skips is used to model all the prosody-dependent allophones. The acoustic-phonetic observation PDF $p(X_{ij}|s_{ij})$ in (10) is modeled as a 3-component mixture Gaussian, 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 created by eliminating some of the low-frequency function-word-dependent phonemes in the SPHINX phoneme set [30]. A 32 dimensional acoustic-phonetic feature vector consists of 15 MFCC coefficients, energy, and their deltas. The one dimensional acoustic-prosodic feature described in Section III.B is modeled as a separate stream in HTK.

The prosodically labeled data consist of 300 utterances (about 3 h of speech sampled at 16 Khz) read by five professional announcers (3 female, 2 male) containing a vocabulary of 3777 words. Training and test sets are formed by randomly selecting 90% of the utterances for training and the remaining 10% for testing. Ideally, the training and test sets should be selected in a speaker-independent fashion (i.e., training and test set should not contain utterances from the same speakers) in order to avoid speaker dependent effects. However, the small size of the database makes speaker-independent experiments impractical, thus we have opted for experiments that include data from all five speakers in both training and test sets.

To measure the strength of the prosody induced allophonic variation modeled in the acoustic model $p(O|Q, H)$, we conducted prosody-dependent allophone recognition experiments on the Radio News Corpus. The reference prosody dependent allophone transcriptions were created by combining the hand-transcribed phonetic transcriptions with the ToBI prosody transcriptions. Two sets of allophone models were constructed: a prosody-dependent set PD, created by splitting each monophone in the SPHINX set into four allophones that implement a four-way prosodic distinction (neutral, lengthened, accented, and accented+lengthened), and a baseline prosody-independent set PI, created in the similar way as does PD except that the prosodic distinctions are implemented logically, i.e., the duration PDFs split from the same phonetic state are tied regardless of the prosodic context and the acoustic-prosodic observation PDFs are removed. Although PI contains the same number of allophones as does PD, it can not detect any acoustic-prosodic effects from the signal. The difference between the allophone recognition correctness and accuracy of PD and PI models when null grammar is used, as given in Table V, reflects the degree of prosody-dependent allophonic variation that is implemented in the PD model. Table V shows that modeling prosody dependence greatly improves the allophone recognition correctness and accuracy with only a small increase in parameter count.

TABLE V
Percent Allophone Recognition Correctness and Accuracy on Prosody Dependent Allophones, and Number of Parameters of the Allophone Models. Both PI and PD Contain 204 Allophones

	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

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 8 variants realizing an eight-way prosody dependent distinction. We found the entropy of the test text increased from $H(W) = 7.02$ bits to $H(W, P) = 8.41$ bits after prosody dependence is implemented; the number of parameters in the language model increased from 5380 to 14 751. The fact that the cross-entropy did not increase by 3 bits suggests that there is a strong correlation between prosody and word strings: quite a large number of words in the test set are uttered with the same prosodic pattern as they were in the training set. By construction, this database includes many word string repetitions, thus word strings in the training data often re-appear in the test data with the same prosodic pattern. $H(W, P)$ can be made comparable with $H(W)$ by marginalizing over all possible prosody sequence P. This marginalization results in an entropy of 5.91 bits, 1.11 bits smaller than $H(W)$ [37]. Consistently with this entropy reduction, we found that with the same acoustic model (PI), the language model PD can improve word recognition by about 0.6% over the language model PI, as shown in Table VI. After switching to acoustic model PD, the word recognition 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 prosodically plausible and reduces the likelihood of the word hypotheses that are prosodically invalid. This statement is supported by Table VI that shows the word error rate (WER)

TABLE VI
PERCENT WORD ERROR RATE AND PERCENT ACCENT AND INTONATIONAL PHRASE BOUNDARY (IPB) ERROR RATE USING PI AND PD ACOUSTIC MODELS IN COMBINATION WITH PI AND PD LANGUAGE MODELS

	AM	LM	HMM	EDHMM
Word	PI	PI	25.11	24.85
	PI	PD	24.48	24.33
	PD	PD	23.50	23.38
Accent	PI	PI	44.59	44.63
	PI	PD	23.25	23.08
	PD	PD	20.39	20.35
IPB	PI	PI	15.53	15.57
	PI	PD	14.67	14.47
	PD	PD	14.51	14.35

TABLE VII
PHONEME RECOGNITION ACCURACY (%) OF THE ALLOPHONE SETS PPI AND PPD WITH 750, 1000, AND 1250 DISTINCT HMM STATES

	750	1000	1250
PPI	39.00	38.61	38.44
PPD	39.14	39.83	38.77

of PD+PD+EDHMM has been reduced by about 1.8% absolute (6.9% relative) against the baseline system PI+PI+HMM.

B. Prosody-Dependent MFCC Distinctions

The second set of experiments used hidden Markov models to represent only prosody-dependent distinctions in the acoustic-phonetic observation PDF $p(X|Q,H)$. Two sets of experiments were conducted. One set of experiments used the tree-clustered allophone models constructed in the tree-based prosody-dependent experiment (TPD) reported in Section III.C. The second experiment used the models constructed manually, based on a log-likelihood improvement criterion, in the triphone-compensated log-likelihood experiment (TLL) reported in Section III.C.

Both experiments TPD and TLL use standard three-state hidden Markov models, with no explicit duration model, and no acoustic-prosodic observation PDF. In both experiments, the total number of HMM states is controlled, so that the prosody-dependent (PD) and prosody-independent (PI) recognizers are allowed to have approximately the same number of total recognizer parameters. In both experiments, allophones in the PI recognizer are distinguished exclusively on the basis of phoneme-context questions, while allophones in the PD recognizer may be distinguished on the basis of either phoneme context or prosodic context; thus the questions available during the design phase of the PD recognizer are a strict superset of the questions available during the design phase of the PI recognizer. The two experiments differed in two important respects. First, the TLL system used a larger selection of prosodic context questions. Second, the TPD system allows context-dependent splitting and tying of individual HMM states, while the TLL system allows splitting only of complete 3-state allophone models. In fact, the TLL system is constrained to use exactly three allophones for every SPHINX monophone.

TABLE VIII
WORD ERROR RATE (%) OF THE PROSODY-INDEPENDENT AND PROSODY-DEPENDENT RECOGNIZERS DEVELOPED IN THE TRIPHONE-COMPENSATED LOG-LIKELIHOOD-BASED (TLL) EXPERIMENT. EACH RECOGNIZER USES A TOTAL OF 138 ALLOPHONES: THREE ALLOPHONES OF EACH PHONEME

Context Definition	Word Error Rate
Triphone Context Only	36.2%
Triphone and Prosodic Contexts	25.4%

Table VII lists phoneme recognition accuracy of the tree-clustered HMM allophone model sets PPI (created from TPI) and PPD (created from TPD) at three levels of recognizer complexity. The three levels of recognizer complexity are determined by the total number of HMM states in the recognizer, including all allophone models: the three levels of complexity are characterized by 750, 1000, and 1250 total HMM states, respectively. For a given level of recognizer complexity the PPD allophone set has better phoneme recognition accuracy than the PPI allophone set.

Table VIII lists word error rate of the triphone-compensated HMM allophone model sets. Because this experiment required a far smaller number of decision trees than the TPD experiment (one per phone instead of one per state), it was possible to train the recognizer on the small ToBI-labeled part of the Radio News Corpus, and therefore it was possible to use a much wider variety of prosodic context questions than the set of questions considered in Table VII. In this experiment, allowing the use of prosodic context in the definition of an allophone model results in an 11% absolute reduction in word error rate.

V. CONCLUSION

The phonetic and linguistic literature on prosody suggests two hypotheses about the interaction between prosody and word recognition. First, prosodic and phonemic context jointly influence the duration, the pitch and the short-time spectrum of a phoneme. Second, prosody is constrained by word strings (and vice versa). In this paper, a prosody dependent speech recognizer that models word and prosody in a unified probabilistic framework is proposed to test if modeling prosody as hidden variables in an HMM based speech recognizer would improve word recognition. Our analyzes and experiments indicate that explicit models of prosody only yield statistically significant reductions in word error rate if the prosodic variables are constrained by explicit models of both acoustic-prosodic interactions and prosodic-language interactions.

Our approach is motivated by an information-theoretic analysis, showing that prosody dependent recognition can increase the mutual information between true word hypothesis and acoustic observation by utilizing the interaction between the acoustic model and the language model of a speech recognizer. The influence of prosody on phonetic distribution is investigated experimentally. Experiments demonstrate that prosody (the intonational phrase boundary and the pitch accent) significantly affects the duration and the pitch of all tested allophones. Modeling of prosody-dependent influences on the short-time spectrum (e.g., MFCC) is much more difficult. Experiments certainly rule out any acoustic distinction between

accented and unaccented phrase-final consonants, or between accented and unaccented phrase-initial consonants; this finding seems to correspond with the strengthened versus lengthened distinction proposed in earlier studies of speech articulation. In comparisons with comparable parameter counts for both prosody-dependent and prosody-independent recognizers, other prosodic distinctions seem to be useful only when they are available as questions to be asked by an allophone clustering algorithm; prosodic questions are selected more or less frequently by the clustering algorithm depending on the variety of prosodic and phonetic questions available to the clustering algorithm.

To accurately model phoneme duration, we implement the recognizer using the explicit duration hidden Markov model (EDHMM) and compared it with systems using HMM. We find that explicit phoneme duration PDFs are far more precise (lower in entropy) if prosodic context is taken into consideration. To model pitch accent, a new acoustic prosodic feature, generated by an ANN from normalized pitch, is incorporated into the acoustic observation, and is modeled by a single Gaussian PDF. In word and prosody recognition experiment on the Radio News Corpus, we find that the proposed prosody dependent recognizers reduce word error rate by as much as 11% over prosody independent recognizers with comparable parameter count.

Our research clearly demonstrated that linguistically well-attested prosodic phenomena (such as pre-boundary lengthening and pitch accents) can be modeled in HMM based speech recognizers to improve word recognition performance with careful feature normalization and model parameterization.

APPENDIX
TRAINING AND DECODING ALGORITHMS FOR EXPLICIT DURATION HMM

In order to provide sufficient background for those without access to Ferguson's paper [28], (19)–(35) and (38)–(41) review Ferguson's EDHMM training algorithm. Equations (36)–(37) and (42)–(45) describe our extensions of Ferguson's algorithm, created for the purpose of applying this algorithm in contemporary continuous speech recognition.

In [28], the forward probabilities (α, α^*) and backward probabilities (β, β^*) are defined as follows:

$$\alpha_t(i) = Pr\{O_1O_2\dots O_t \text{ and state } i \text{ ends at } t\} \tag{19}$$

$$\alpha_t^*(i) = Pr\{O_1O_2\dots O_t \text{ and state } i \text{ starts at } t+1\} \tag{20}$$

$$\beta_t(i) = Pr\{O_{t+1}O_{t+2}\dots O_T, \text{ given that state } i \text{ ends at } t\} \tag{21}$$

$$\beta_t^*(i) = Pr\{O_{t+1}O_{t+2}\dots O_T, \text{ given that state } i \text{ starts at } t+1\} \tag{22}$$

where t is the time index, i is the state index, and O_t is the observation vector at time t. The forward and backward probabilities can be computed recursively with proper initialization

$$\alpha_t^*(j) = \sum_{i=1}^{N} \alpha_t(i)a(j|i) \tag{23}$$

$$\alpha_t = \sum_{\tau<t} \alpha_{t-\tau}^*(i)d(\tau|i)b(O_{t-\tau+1}\dots O_t|i) \tag{24}$$

$$\beta_t(i) = \sum_{j=1}^{N} a(j|i)\beta_t^*(j) \tag{25}$$

$$\beta_t^*(j) = \sum_{\tau<t} \beta_{t+\tau}(i)d(\tau|i)b(O_{t+1}\dots O_{t+\tau}|i) \tag{26}$$

where $a(j|i)$ is the transition probability from state i to state j, $d(\tau|i)$ is the probability of staying in state i with duration τ, and $b(O_1O_2\dots O_t|i)$ is the probability of observing $O_1O_2\dots O_t$ in state i. Under these definitions, $S_t(i)$, the expected number of times state i started at time t or before, and $E_t(i)$, the expected number of times state i terminated before time t, can be computed

$$S_t(i) = \sum_{\tau<t} \alpha_\tau^*(i)\beta_\tau^*(i)/P, \tag{27}$$

$$E_t(i) = \sum_{\tau<t} \alpha_\tau(i)\beta_\tau(i)/P \tag{28}$$

where P is the probability of observing $O_1O_2\dots O_T$ given the current set of model parameters. The state residence probability $\gamma_t(i)$, the probability that state i is used to produce O_t, can be computed as

$$\gamma_t(i) = S_t(i) - E_t(i). \tag{29}$$

The following statistics are accumulated across all examples of the phonetic models under re-estimation:

$$C(i,j) = \sum_{r=1}^{R}\sum_{t=1}^{T} \frac{\alpha_t(i)a(j|i)\beta_t^*(j)}{P_r} \tag{30}$$

$$C(i,\tau) = \sum_{r=1}^{R}\sum_{t=1}^{T} \frac{\alpha_t^*(i)d(\tau|i)b(O_{t+1}\dots O_{t+\tau}|i)\beta_{t+\tau}(i)}{P_r} \tag{31}$$

where r is the index of the examples and R is the total number of examples. The transition probability $a(j|i)$, the duration probability $d(\tau|i)$, the mean μ_i and covariance matrix Σ_i of the ith state can be re-estimated as

$$\hat{a}(j|i) = \frac{C(i,j)}{\sum_{j=1}^{N} C(i,j)} \tag{32}$$

$$\hat{d}(\tau|i) = \frac{C(i,\tau)}{\sum_{\tau=1}^{D} C(i,\tau)} \tag{33}$$

$$\hat{\mu}_i = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_i^r(t)O_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_i^r(t)} \tag{34}$$

$$\hat{\Sigma}_i = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_i^r(t)\left(O_t^r - \hat{\mu}_i\right)\left(O_t^r - \hat{\mu}_i\right)'}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_i^r(t)}. \tag{35}$$

When the observation probability distribution is modeled by a mixture Gaussian with M components, the residence probability of the kth mixture component can be computed using

$$\gamma_t^r(j,k) = \gamma_t^r(j)\frac{c_{jk}\mathcal{N}\left(O_t^r, \mu_{jk}, \Sigma_{jk}\right)}{\sum_{m=1}^{M} c_{jk}\mathcal{N}\left(O_t^r, \mu_{jk}, \Sigma_{jk}\right)} \tag{36}$$

where c_{jk} is the weight of the kth mixture component of state j and can be reestimated as

$$\hat{c}_{jk} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} \gamma_t^r(j,k)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{k=1}^{M} \gamma_t^r(j,k)}. \tag{37}$$

The decoding algorithm of EDHMM has a form that is slightly different from the standard Viterbi algorithm due to the nature of the semi-Markov chain. In analogy to forward and backward probabilities, the maximum a posteriori probabilities can be defined as

$$\delta_t(i) = \max \Pr(O_1 O_2 \dots O_t \text{ and state } i \text{ ends at time } t) \quad (38)$$

$$\delta_t^*(i) = \max \Pr(O_1 O_2 \dots O_t \text{ and state } i \text{ starts at time } t+1). \quad (39)$$

Similar to the forward and backward algorithm, the maximum a posteriori probabilities can be computed recursively

$$\delta_t^*(j) = \max_i \delta_t(i) a(j|i) \quad (40)$$

$$\delta_t(i) = \max_\tau \delta_{t-\tau}^*(i) d(\tau|i) b(O_{t-\tau+1} \dots O_t|i). \quad (41)$$

For backtracking, all the arguments that maximize (40) and (41) need to be stored in memory. Define

$$\psi_t(j) = \arg\max_i \delta_t(i) a(j|i) \quad (42)$$

$$\zeta_t(i) = \arg\max_\tau \delta_{t-\tau}^*(i) d(\tau|i) b(O_{t-\tau+1} \dots O_t|i) \quad (43)$$

where $\psi_t(j)$ stores the most likely state i from which the transition into state j occurs, and $\zeta_t(i)$ stores the most likely duration of state i. To recover the best path, the following recursion can be used starting from $t = T$

$$\tau_t^* = \zeta_t(q_t^*) \quad (44)$$

$$q_{t-\tau_t^*}^* = \psi_{t-\tau_t^*}(q_t^*). \quad (45)$$

The Viterbi algorithm introduced above requires $(D+N)/N$ times more operations than does the standard Viterbi algorithm, provided that all the arguments required in (41) are stored in the memory.

REFERENCES

[1] L. Hahn, "Native Speakers," Ph.D. dissertation, 1999. Reactions to nonnative stress in English discourse.

[2] J. F. Pitrelli, M. E. Beckman, and J. Hirschberg, "Evaluation of prosodic transcription labeling reliability in the TOBI framework," in *Proc. ICSLP*, 1994.

[3] M. E. Beckman, "The parsing of prosody," *Lang. Cogn. Processes*, vol. 11, no. 1, pp. 17–67, 1996.

[4] T. Cho, "Effects of prosody on articulation in English," Ph.D. dissertation, Univ. Calif., Los Angeles, 2001.

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

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

[7] 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. Int. Conf. Phonetic Sciences*, 2003.

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

[9] 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, U.K.: Cambridge Univ. Press, 1990, pp. 152–178.

[10] L. Dilley, S. Shattuck-Hufnagel, and M. Ostendorf, "Glottalization of word-initial vowels as a function of prosodic structure," *J. Phonetics*, vol. 24, pp. 423–444, 1996.

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

[12] J. H. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. Eurospeech*, 2001.

[13] P. Taylor, S. King, S. Isard, H. Wright, and J. Kowtko, "Using intonation to constrain language models in speech recognition," in *Proc. Eurospeech*, 1997.

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

[15] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1707–1717, Mar. 1992.

[16] M. E. Beckman and G. M. Ayers, Guidelines for ToBI Labeling: The Very Experimental HTML Version, 1994.

[17] E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in automatic speech processing," in *Proc. ISCA Int. Conf. Speech Prosody*, Nara, Japan, 2004.

[18] D. Vergyri, A. Stolcke, V. Gadde, L. Ferrer, and E. Shriberg, "Prosodic knowledge sources for automatic speech recognition," in *Proc. ICASSP*, 2003.

[19] A. Stolcke, E. Shriberg, D. Hakkani-Tur, and G. Tur, "Modeling the prosody of hidden events for improved word recognition," in *Proc. EUROSPEECH*, 1999.

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

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

[22] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *J. Acoust. Soc. Amer.*, vol. 107, no. 3, pp. 1697–1714, 2000.

[23] S. S. Kim, M. Hasegawa-Johnson, and K. Chen, "Automatic recognition of pitch movements using multilayer perceptron and time-delay recursive neural network," *IEEE Signal Process. Lett.*, vol. 11, no. 7, pp. 645–648, Jul. 2004.

[24] 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,", Tech. Rep., 1996. CSLU 1996 Summer Workshop.

[25] Y. Normandin, "Optimal splitting of hmm gaussian mixture components with mmie training," in *Proc. ICASSP*, vol. 1, 1995, pp. 449–452.

[26] R. Kompe, *Prosody in Speech Understanding Systems*. New York: Springer-Verlag, 1997.

[27] C. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 469–481, Oct. 1994.

[28] J. D. Ferguson, "Variable duration models for speech," in *Proc. Symp. Application of Hidden Markov Models to Text and Speech*, Princeton, NJ, 1980, pp. 143–179.

[29] S. E. Levinson, "Continuously variable duration hidden Markov models for automatic speech recognition," *Comput. Speech Lang.*, vol. 1, no. 1, pp. 29–45, 1986.

[30] K. F. Lee, "Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 38, no. 4, pp. 599–609, Apr. 1990.

[31] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICASSP*, 1998.

[32] K. Chen, M. Hasegawa-Johnson, A. Cohen, and J. Cole, "A maximum likelihood prosody recognizer," in *Proc. ISCA Int. Conf. Speech Prosody*, Nara, Japan, 2004.

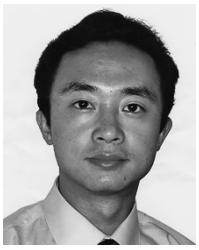
[33] I. Shafran, M. Osendorf, and R. Wright, "Prosody and phonetic variability: Lessons learned from acoustic model clustering," in *ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 127–131.

[34] J. J. Odell, P. C. Woodland, and S. J. Young, "Tree-based state clustering for large vocabulary speech recognition," in *Proc. Int. Symp. Speech, Image Processing and Neural Networks*, 1994, pp. 690–693.

[35] G. Schwartz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 5, no. 2, pp. 461–464, 1978.

[36] S. Borys, "The importance of prosodic factors in phoneme modeling with applications to speech recognition," in *Proc. HLT-NAACL*, Edmonton, AB, Canada, 2003.

[37] K. Chen and M. Hasegawa-Johnson, "Improving the robustness of prosody dependent language modeling based on prosody syntax dependence," in *Proc. IEEE ASRU*, 2003.



Ken Chen received the B.S. degree in precision instruments from the Tsinghua University, Beijing, China, in 1996, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2001 and 2004, respectively.

He is currently a postdoctoral Research Associate in the Department of Chemistry and Biochemistry, University of California, San Diego. His research interests include machine learning, speech recognition, signal processing, and computational biology. He has authored and coauthored 15 papers in professional conferences and journals.

Dr. Chen is a member of Phi Kappa Phi.

Mark Hasegawa-Johnson (M'97–SM'04) received the S.B., S.M., and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 1989, 1989, and 1996, respectively.

He has held engineering internships in echo cancellation at Motorola Labs., Schaumburg, IL, and in speech coding at Fujitsu Laboratories Limited, Kawasaki, Japan. From 1996 to 1997, he was the 19th Annual ASA F.V. Hunt Post-Doctoral Research Fellow, designing and testing models of articulatory motor planning at the University of California at Los Angeles (UCLA) and at MIT. From 1998 to 1999, he held an NIH Individual National Research Service Award at UCLA. In 1999, he became an Assistant Professor of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign, where he co-founded the Illinois Speech and Language Engineering group. He is the author or co-author of four U.S. patents, eight Japanese patents, nine refereed journal articles, more than 50 conference papers, and a chapter in the *Encyclopedia of Telecommunications* (New York: Wiley).

Dr. Hasegawa-Johnson is Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. He is a member of the IEEE Signal Processing Society, the Acoustical Society of America, the Audio Engineering Society, and the International Speech Communication Association. He is also a member of Eta Kappa Nu, Tau Beta Pi, Sigma Xi, and Phi Beta Kappa, and has been listed in *Marquis Who's Who in Science and Technology*.

Aaron Cohen received the B.S. degree in computer engineering and the M.S. degree in electrical engineering from the University of Illinois at Urbana-Champaign (UIUC) in 2002 and 2004, respectively.

From 2000 to 2001, he was an Intern at the Allstate Corporation. From 2002 to 2004, he was a graduate Research Assistant in the Illinois Speech and Language Engineering Group, where he used neural networks and classification trees to study the relationship between prosody and syntax. He is the author or coauthor of four papers in professional conferences and journals.

Mr. Cohen is a member of Eta Kappa Nu and Tau Beta Pi.

Sarah Borys received the B.S. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign in 2001 where she is currently pursuing the M.S. degree.

From 2002 until 2003, she was an undergraduate Research Assistant in the Illinois Speech and Language Engineering Group. In 2003, she became a graduate Research Assistant in the same group. She is the author or coauthor of seven papers in professional conferences and journals. Her current research uses machine learning methods, such as hidden Markov models and support vector machines, to model the acoustic correlates of prosody, of distinctive features, and of disfluency, for the purpose of improved automatic speech recognition.

Sung-Suk Kim received the B.S. degree in electrical engineering from Yeungnam University, Gyeongsan, Korea, in 1985, and the M.S. and Ph.D. degrees in electronics and computer engineering from the University of Ulsan, Ulsan, Korea, in 1987 and 1990, respectively.

From 1985 to 1991, he was with Korea Electric Power Corporation (KEPCO). He was a Visiting Professor at Beckman Institute, University of Illinois at Urbana-Champaign, in 2003. He is currently an Associate Professor with the School of Computer and Information, Yong-In University, Yongin, Korea. His research interests includes speech recognition, computer-assisted language learning (CALL), soft computing, and blind source separation.

Jennifer S. Cole received the B.A. and M.A. degrees in linguistics from the University of Michigan, Ann Arbor, in 1983 and 1984, respectively, and the Ph.D. degree in linguistics from the Massachusetts Institute of Technology, Cambridge, in 1987.

She has been a faculty member in Linguistics and Computer Science at the University of Illinois at Urbana-Champaign since 1989, as visiting Assistant Professor (1989–1990), Assistant Professor (1990–1998), and Associate Professor (1998–present). She has been a member of the Cognitive Science Group at the Beckman Institute on the University of Illinois campus since 1990. She served as an Instructor in linguistics at Yale University in 1987–1989. She is the author of one book, seven refereed articles, and more than 20 conference papers, and is editor of three books in the areas of phonological theory, laboratory phonology, and acoustic phonetics. She has published in the *Handbook of Phonological Theory* (Cambridge, MA: Blackwell) and is the author of three encyclopedia articles on the Sindhi language. She has served on the editorial boards of the *Linguistic Inquiry* and *Phonology*, and chaired the 9th Conference on Laboratory Phonology. She has received research support from the NSF, the NIH, the Department of Education, the NSEP, and NASA.

Dr. Cole received several research fellowships from the University of Illinois, and was an IBM Graduate Student Fellow, a National Science Foundation Graduate Fellow, and an Ida M. Green Scholar (MIT). She is a member of the Linguistic Society of America, the International Phonetic Society, and Phi Beta Kappa.

Jeung-Yoon Choi received the B.Sc. and M.Sc. degrees in electronic engineering in 1992 and 1994, respectively, from Yonsei University, Seoul, Korea, and the Ph.D. degree in electrical engineer and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, in 1999.

She was engaged in postdoctoral research at MIT from 1999 to 2001. She has been with the University of Illinois at Urbana-Champaign since 2002. Her interests are in speech communication, signal processing, and acoustics.