Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 46 (2005) 418–439
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus ☆

Mark Hasegawa-Johnson *, Ken Chen, Jennifer Cole, Sarah Borys, Sung-Suk Kim, Aaron Cohen, Tong Zhang, Jeung-Yoon Choi, Heejin Kim, Taejin Yoon, Sandra Chavarria

Beckman Institute, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, United States

Received 1 September 2004; received in revised form 17 December 2004; accepted 23 January 2005

Abstract

This paper describes automatic speech recognition systems that satisfy two technological objectives. First, we seek to improve the automatic labeling of prosody, in order to aid future research in automatic speech understanding. Second, we seek to apply statistical speech recognition models of prosody for the purpose of reducing the word error rate of an automatic speech recognizer. The systems described in this paper are variants of a core dynamic Bayesian network model, in which the key hidden variables are the word, the prosodic tag sequence, and the prosody-dependent allophones. Statistical models of the interaction among words and prosodic tags are trained using the Boston University Radio Speech Corpus, a database annotated using the tones and break indices (ToBI) prosodic annotation system. This paper presents both theoretical and empirical results in support of the conclusion that a *prosody-dependent* speech recognizer—a recognizer that simultaneously computes the most-probable word labels and prosodic tags—can provide lower word recognition error rates than a standard prosody-independent speech recognizer in a multi-speaker speaker-dependent speech recognition task on radio speech.

Keywords: Automatic speech recognition; Prosody

1. Introduction

Prosody (the rhythm and intonation patterns of spoken language) helps listeners to understand speech with minimum cognitive load (Hahn, 1999). The acoustic cues of duration, intensity and pitch combine to confer prosodic prominence

☆ Supported by NSF award number 0132900, and by a grant from the University of Illinois. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF or the University of Illinois.
* Corresponding author.
E-mail address: jhasegaw@uiuc.edu (M. Hasegawa-Johnson).

0167-6393/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.specom.2005.01.009

or stress at two levels in English: lexical stress is located on one syllable within a word, and phrasal prominence is located on one or more words within a phrase. Lexical stress provides a partial indicator of word boundaries in continuous speech, and phrasal prominence plays an important role in signaling the contribution of a word to the evolving discourse. Lexical stress and phrasal prominence are relevant to the development of speech recognition systems because they each affect the pronunciation of the individual phonemes that make up the words in an utterance. The consonants and vowels in a prominent syllable are pronounced with greater clarity and duration than in a syllable without prominence. In English, these effects are especially dramatic: in syllables lacking lexical stress, consonants and vowels are significantly reduced in both time and frequency dimensions (Kent and Netsell, 1971), leading to ambiguity in segmentation of the speech stream and in identification of the individual phonemes. In addition to marking prominence, prosodic cues also serve to signal the phrase structure of the utterance, identifying word sequences that must be grouped together in the construction of the sentence meaning in a given discourse context. Prosodic phrasing is also relevant to speech recognition because the intonational and rhythmic patterns that signal major phrase breaks and sentence boundaries provide acoustic analogs to punctuation marks and font-based highlighting present in text.

The research described in this paper has two goals. First, we seek to improve the automatic labeling of prosody, in order to aid future research in automatic speech understanding. Second, we seek to apply statistical speech recognition models of prosody for the purpose of reducing the word error rate of an automatic speech recognizer. The proposed models are built around a core dynamic Bayesian network model, in which the key hidden variables are the word, the prosodic tag sequence, and the prosody-dependent allophones. In the task of word recognition, the prosodic labels may be known, unknown, or ignored. In the task of automatic prosody annotation, the word labels may be known, unknown, or ignored. In the typical situation confronted by an automatic speech recognizer, neither the prosodic tag sequence nor the word labels are known a priori. This paper presents both an information-theoretic analysis and a large number of statistical results in support of the conclusion that a *prosody-dependent* speech recognizer—a recognizer that simultaneously computes the most-probable word labels and prosodic tags—can provide lower word recognition error rates than a standard prosody-independent speech recognizer.

Many results described in this paper have been previously published in short papers, conference papers, and theses, including (Chen et al., 2003a; Chen et al., 2003b; Chen and Hasegawa-Johnson, 2003; Chen and Hasegawa-Johnson, 2004; Chen et al., 2004b; Chen et al., 2004a; Hasegawa-Johnson et al., 2004; Kim et al., 2004b; Cole et al., 2003; Kim et al., 2004a; Borys, 2003; Cohen, 2004; Ren et al., 2004; Chavarria et al., 2004). A key objective of this paper is to bring these several papers together in one location, in order to present a coherent summary of results to the automatic speech recognition and prosody communities.

2. Background

The task of a speech recognizer, given a sequence of observed short-time slices of the acoustic spectrum, is to find the sequence of word models that maximizes the recognition probability:

$$\begin{aligned}\widehat{W} &= \arg\max p(W|O) \\ &= \arg\max p(O|W)p(W) \\ &\approx \arg\max p(O|Q)p(Q|W)p(W)\end{aligned} \tag{1}$$

where $Q = [q_1, \ldots, q_L]$ is a sequence of sub-word units, typically clustered triphones, $p(O|Q)$ is the acoustic model, $p(Q|W)$ is the pronunciation model, and $p(W)$ is the language model. The systems reported in this paper use a non-probabilistic dictionary, meaning that the search algorithm assumes $p(Q|W) = 1$ for all allowed pronunciations. The dictionary for these experiments was created based on the dictionary supplied with the Radio Speech Corpus. No phonological expansion was applied, thus the only alternate pronunciations in the dictionary are those supplied by human

transcribers; the pronunciation density is 1.01 pronunciations/word.

2.1. Modeling prosody

There are at least three components of prosody that can be incorporated into a speech recognition system: lexical stress, prosodic phrasing, and phrasal prominence. Lexical stress is marked invariantly for each word in its dictionary transcription, designating the primary stressed syllable within the word. Prosodic phrasing groups words in a sentence into a hierarchical structure, and the key word in each phrase receives an extra phrasal prominence on its lexically stressed syllable. Phrasal prominence is cued variably by pitch accent, increased duration, and possibly increased energy, centered on or immediately after the lexically stressed syllable in the prominent word, and marks a word as contributing new information to the discourse, among other functions. For example, phrasal prominence on the words "banana" and "hand" signal the important words in the sentence "He has a baNAna in his HAND". If the sentence is uttered as an answer to the question "What does the gorilla have", the phrasal prominence of the word "banana" may be perceptually stronger than that of the word "hand;" the former word is then said to carry an emphatic or contrastive accent.

Phrasal prominence is usually realized on a lexically stressed syllable, and thus also serves as a cue to lexical stress. Conversely, a syllable without lexical stress is often realized with a non-distinctive, reduced vowel quality (schwa) in English, as in the first syllable of the word "banana". A number of phonological processes affecting consonant realization also depend on lexical stress, such as the process that realizes a /t/ or /d/ as a flapped sound, as in the medial consonant of "butter". Both lexical stress and phrasal prominence cause acoustic changes in consonants and vowels that are large enough to potentially confound a speech recognition system. Lexical stress is deterministic, specified in the dictionary entry for all occurrences of a word, and as such it is not difficult to use in speech recognition. But there is little to be gained by doing so. All standard recognizers explicitly model the difference between reduced vowels (schwa) and non-reduced vowels in dictionary entries, without direct reference to stress (Lee and Hon, 1989; Zue et al., 1990). In the absence of phrasal prominence, stress-related differences other than vowel reduction have been found to be too small to be useful for speech recognition (van Kuijk and Boves, 1999).

Prosodic phrasing and phrasal prominence are not deterministic, but vary depending on the syntax, dialog context, and speaking style of any particular utterance. These aspects of prosody are more difficult to model than lexical stress, but may be more rewarding. The prosodic phrase structure and prominences of an utterance may be coded in a vector of auxiliary word annotation, $P = [p_1, \ldots, p_M]$ of length equal to the word string $W = [w_1, \ldots, w_M]$. For each word w_i in W, the entry p_i in P specifies the depth of the phrase boundary following w_i and the level of phrasal prominence on the word w_i.

For our research we have adopted a reduced form of the ToBI (Tones and Break Indices) notational conventions. ToBI is an international standard for prosodic transcription among speech scientists across disciplines (Silverman et al., 1992), which has been used to annotate at least four English-language databases of recorded speech including the DCIEM Map Task corpus, the Boston University Radio Speech Corpus (Ostendorf et al., 1995), the Boston Directions Corpus (Hirschberg and Nakatani, 1998), and two subsets of the Switchboard Corpus (Chavarria et al., 2004; Ostendorf et al., 2002). The ToBI system marks the strength of the boundary between words in an utterance by a "break index" between 0 and 4, and also marks the tonal melody expressing phrasal prominence. The break indices distinguish the normal word boundary ("level 1") from the weaker boundary that allows a function word to graft onto an adjacent word ("level 0"), e.g. between the words "in" and "the" in the phrase "in the dark". Stronger boundaries separate prosodic phrases at two levels: the intonational ('major') phrase and the intermediate ('minor') phrase. An intonational phrase is built up from one or more intermediate phrases, which in turn may consist of a sequence of one or more

words. The hierarchical prosodic phrase structure constrains the placement of phrasal prominences, which are assigned one per intermediate phrase, with a culminating main prominence on the intermediate phrase that contributes the most important information to the utterance. The prosodic phrase boundaries are partially determined by syntactic phrase structure, but prosodic and syntactic phrases are not isomorphic, and the discrepancy is in part responsible for the challenge of prosodic parsing in speech recognition. Thus, each entry in the vector P consists of a break index (0, 1, 3, or 4) and, if the word has phrasal prominence, a symbol describing the tonal melody of the prominence (e.g., H* + L marking a high-falling pitch contour). Examples of possible prosodic transcriptions of simple sentences are given in Table 1.

The ToBI system labels pitch accent tones, phrase boundary tones, and prosodic phrase break indices. 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, L–H%, representing a medial pitch (sometimes called a "continuation rise"), H–H%, representing a canonical yes–no question contour, and H–L%; the contours !H–L% and !H–H% are less frequently observed. Seven types of accent tones are labeled: H*, !H*, L + H*, L + !H*, L*, L* + H and H + !H*.

The ToBI system has the advantage that it can be used consistently by labelers for a variety of styles. For example, if one allows a level of uncertainty in order to account for differences in labeling style, it can be shown that the different transcribers of the Radio Speech Corpus agree on break index with 95% inter-transcriber agreement (Ostendorf et al., 1995; Pitrelli et al., 1994). 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 L + H* versus L* + H distinction is subject to considerable inter-transcriber disagreement (Ostendorf et al., 1995).

2.2. Acoustic and articulatory correlates of prosody

The phrasal prominences and boundaries we propose to model, described above, are signaled in part by pitch events that can be identified through analysis of the variation in fundamental frequency of the acoustic signal. In addition, prosodic phrase boundaries and phrasal prominence are also marked by changes in the relative timing of the movements of different articulators, which cause a variety of linked changes in the acoustic signal.

First, durations change. Wightman et al. (1992) found that the average normalized duration of phonemes in the rhyme of the syllable preceding a phrase boundary increases monotonically as a function of the depth of the boundary.

Second, syllables with phrasal prominence tend to be produced with higher energy than surrounding syllables. A number of studies have found that, in spontaneous or conversational speech, phrasal prominence is more robustly cued by changes in duration and energy than it is by changes in F0 (e.g., Batliner et al., 1997; Greenberg and

Table 1
Sample prosody of the words "kids play in the park"

Punctuated orthography	Word string	ToBI prosodic tag string
"Kids play in the park".	W = (kids, play, in, the, park)	P = (3 H*L–, 1, 0, 1, 4 H*L–L%)
"Kids play, in the park".	W = (kids, play, in, the, park)	P = (1, 3 H*L–, 0, 1, 4 H*H–L%)
"Kids, play in the park!"	W = (kids, play, in, the, park)	P = (4 H*L–H%, 3 H*, 0, 1, 4 H*L–L%)

Prosody in the first row would be appropriate in response to the question "Who plays in the park?" Prosody in the second row would be appropriate in response to "What do kids do on Saturday?" Prosodic tags recognized in this paper are a subset of those exemplified in column 3.

Hitchcock, 2001). Casual observations often suggest that energy distinguished lexically stressed from unstressed syllables, but when the dimensions of lexical stress and vowel reduction are separately controlled, there is no difference in energy between stressed and unstressed syllables (van Kuijk and Boves, 1999). Either phrasal prominence or lexical stress may be cued by more subtle acoustic measurements such as the integral of energy over the duration of the syllable (Greenberg and Hitchcock, 2001), or spectral tilt (Sluijter et al., 1997).

Third, changes in articulatory timing may allow consonants to be "more consonantal", in the sense that they have longer closure durations, and in the sense that their closures cover a larger area of the palate (Fougeron and Keating, 1997). Both voiced and unvoiced stop consonants exhibit longer voice onset times in pitch accented than in unaccented syllables (Choi et al., 2003; Cole et al., 2003; Kim et al., 2004a). Vowel productions are also more extreme when the syllable has phrasal prominence, but studies differ on the way in which the extreme production is realized: some studies say that vowels in prominent syllables are more canonical (Hombert, 1978), in the sense that low vowels are lower while high vowels are higher; another study suggests that low vowels and diphthongs are far more likely than high vowels to be perceived as having phrasal prominence (Greenberg and Hitchcock, 2001). Taken together, the results of both consonant and vowel studies suggest that phrasal prominence in English may be implemented as a form of localized hyperarticulation (DeJong, 1995), thus, in some yet-to-be-specified way, consonants are more "consonantal" in a pitch-accented syllable, while vowels are more "vocalic".

2.3. Recognition of prosody

Much successful previous work in the automatic recognition of prosody has relied on an auxiliary matrix of word-level observations, $Y = [y_1, \ldots, y_M]$, where each entry y_i in Y includes information about the pitch contour and phoneme durations during the span of time claimed by the corresponding proposed word w_i. Wightman and Ostendorf (1994) and Kompe (1997) have shown that it is possible to disambiguate sentences with identical word strings but different prosody (e.g. "kids play in the park" vs. "kids, play in the park!") using a criterion of the form

$$\begin{aligned}\widehat{P} &= \arg\max p(X, Y, Q, W, P) \\ &= \arg\max p(X|Q)p(Q|W)p(Y|P)p(P|W)p(W)\end{aligned} \tag{2}$$

Kompe implements the probability $p(Y|P)$ using a neural network trained to estimate the probability mass function $p(P|W, Y)$, where

$$p(P|W, Y) = \prod_{i=1}^{M} p(p_i|w_{i-2}, w_{i-1}, w_i, y_i) \tag{3}$$

Kompe found that when a proposed word string is correct, the most likely prosody almost always has a probability $p(P|W, Y)$ very close to 1.0. By eliminating from consideration any word string without a likely accompanying prosody, Kompe was able to improve the computational efficiency of a large-vocabulary telephone-based dialog system by a factor of two or three with only small decreases in the word recognition accuracy.

To build 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 (Beckman and Elam, 1994). Shriberg and co-workers (Shriberg and Stolcke, 2004; Vergyri et al., 2003; Stolcke et al., 1999; Ferrer et al., 2003) have proposed a different approach that makes use of acoustic prosodic cues without requiring explicit prosodic labeling. In their approach, acoustic prosodic cues (e.g., pitch, energy) are conditioned over a set of hidden event variables of interest for rich text transcription, including sentence and topic boundaries, disfluency markers, dialog act labels, and talker identity. The hidden variables are also usually conditioned on the recognized word string (or on other prior knowledge, e.g., knowledge of the frequency with which talker turn changes occur). By combining information from the recognized word string and from acoustic prosodic cues, it is possible to obtain high-accuracy estimates of the hidden event variables (Liu et al., 2003; Shriberg and Stolcke, 2004).

Taylor (2000) demonstrates one of the few previous systems able to recognize pitch accents without prior information about word boundary location. His two-stage prosody recognition system first locates pitch events using an HMM, then labels the pitch events using an analysis-by-synthesis matching strategy. The best reported pitch event recognition system uses three-state mixture-Gaussian hidden Markov models of each distinct pitch event label; there are 13 distinct pitch event labels, including silence and all possible combinations of four accent levels with three boundary types. The HMM observes speaker-normalized F0 and delta-F0. Results are scored using the HTK HResults program, modified so that a recognized event is considered correct only if it covers at least 50% of the observation frames covered by a true pitch event. Under these constraints, speaker-independent pitch event recognition correctness on one talker from the Boston University Radio Speech Corpus is 72.7%, with a recognition accuracy of 47.7% (25% insertion rate).

Wightman and Ostendorf (1994) and Ostendorf and Ross (1997) considered the problem of automatically locating pitch accents and intonational phrase boundaries, given knowledge of syllable start times and end times, but with limited explicit use of word sequence information. In their system, decision-tree models were trained to calculate the posterior probability of syllable-level prosody labels given the syllable-timed acoustic features; Wightman and Ostendorf also made use of N-gram sequence information. Using these methods, Wightman and Ostendorf were able to correctly label the pitch accent status of 84% of words in an arbitrarily selected test subset of the Radio Speech Corpus, while Ostendorf and Ross were able to correctly label 89% of syllables (chance performance in the first task is about 55%, and the inter-transcriber agreement rate is 95%; chance levels are somewhat higher in task of Ostendorf and Ross). Wightman and Ostendorf also considered the problem of intonational phrase boundary (IPB) detection, but their system performed poorly on this task: IPB recognition rate was only 71%, 13% below the chance level of 84%. Their results confirm that acoustic information alone is insufficient to automatically label intonational phrase boundary position.

Cohen (2004) considered a problem complementary to that of Wightman and Ostendorf: the problem of recognizing IP boundaries and pitch accents based exclusively on the word sequence, with no information at all from the acoustic waveform. He compared six types of learning algorithms: three types of tree-based learners, a boosting algorithm, a neural network with raw indicator variables at the input, and a neural network with heuristically clustered input variables. Using data from the Boston University Radio Speech Corpus, both IP boundary location and pitch accent location were most accurately predicted using the neural network with heuristically clustered input variables: IP boundary location was recognized with 89.6% accuracy, while pitch accent was recognized with 83.1% accuracy. Both systems performed well above chance, and well above the performance of acoustic-only prosody recognizers, but also somewhat below the reported inter-transcriber agreement rates. Cohen suggested that, in order to achieve the best possible automatic labeling of IPB and pitch accent, a system should use input information extracted from both the syntactic features of the word string and the acoustic-prosodic features of the waveform.

2.4. The Boston University Radio Speech Corpus

Experiments reported in this paper make use of the Boston University Radio Speech Corpus, one of the largest speech corpora published with manual prosodic annotations intended for speech technology experiments (Ostendorf et al., 1995). The Radio Speech Corpus consists of recordings of broadcast radio news stories including original radio broadcasts and laboratory broadcast simulations recorded from seven FM radio announcers (4 male, 3 female). Radio announcers usually use more clear and consistent prosodic patterns than non-professional readers, thus the Radio Speech Corpus comprises speech with a *natural but controlled* style, combining the advantages of both read speech and spontaneous speech.

All paragraphs in this corpus are annotated with the orthographic transcription, automatically

generated part-of-speech tags, and automatically generated phone alignments. A large number of paragraphs are also annotated with prosodic labels. The prosodic labeling system represents prosodic phrasing, phrasal prominence and boundary tones, using the ToBI system for American English (Beckman and Elam, 1994). In addition to the canonical 7 pitch accent types, the notation "*?" is used to mark a questionable pitch accent (an accent whose type the transcriber could not discern). Some transcriptions mark the location of a pitch accent (as "*") but not its type; most of these seem to be high or downstepped accents.

The Radio Speech Corpus is one of the few prosodically transcribed corpora for which inter-transcriber agreement statistics are available (Pitrelli et al., 1994; Ostendorf et al., 1995). For example, agreement between transcribers on the presence vs. absence of pitch accent is approximately 91%. Agreement on the exact type of pitch accent is somewhat lower. The vast majority of pitch accents in the Radio Speech Corpus are centered on a high pitch movement (71%) or a downstepped pitch movement (25%). Inter-transcriber agreement on the level of phrase break separating two words is 95%. Inter-transcriber agreement rates on the type of phrase boundary tone have not been reported for the Radio Speech Corpus, but our experience transcribing Switchboard (Yoon et al., 2004; Godfrey et al., 1992) suggests that, given the level of break index, boundary-tone disagreements are rare.

Many paragraphs in the Radio Speech Corpus are prosodically transcribed, but not all. The total duration of prosodically transcribed data is 3.5 h. Transcriptions are not uniformly distributed among the seven talkers. Nearly half of the prosodically transcribed speech data (about 90 min) are from one female talker (F2B). Twenty to 30 min of transcribed data are available for each of four other talkers (F1A, F3A, M1B, and M2B); very little transcribed data are available for the other two talkers. In order to train automatic speech recognition systems using these data, two different methods were used to partition the data: a "speaker-independent" method for the training of prosody detection algorithms, and a "multi-speaker speaker-dependent" partition for the training of speech-to-text algorithms. These training/test partitions are summarized in Table 2.

The prosodic event detectors described in Sections 5 and 8 are implemented using classifiers with a relatively low parameter count: the neural networks in Section 5 each have about 700 trainable parameters, and the hybrid mixture-Gaussian + neural-network classifier in Section 8 has about 4000 trainable parameters. Using an old heuristic rule (the number of training tokens per class should be five times the number of trainable parameters), it is possible to estimate that the classifier in Section 5 should be trained using 3500 training tokens per class, while the system in Section 8 should be trained using 20,000 tokens per class. It is possible to obtain nearly sufficient training corpora for these classifiers using speaker-independent partitions of the Radio News Corpus. The pitch accent detector in Section 5 was trained using the data from talker F1A (2038 pitch accent tokens, or slightly fewer than the recommended number), and tested using the data from talker

Table 2
The experiments reported in this paper used different train/test partitions of the Radio Speech Corpus, depending on the number of trainable parameters of each classifier

Section	Algorithm	Recognized label	Training set	Development test	Test set
5	Neural net	Pitch accent	F1A	F2B	
6	HMM	Phrase boundary	A + B		C
7	HMM	Words, accent, boundary	A	B	C
8	Neural net	Accent, boundary	F2B + 3 talkers	F1A, M1B, or M2B	

Sections 5 and 8 describe speaker-independent prosodic event detectors; for those experiments, training and test corpora are described using talker labels F1A, F2B, M1B, and M2B. Sections 6 and 7 describe multi-speaker speaker-dependent speech recognition systems. Those two sections use the same three-way partition of the database: A = a set with 85% of utterances from all talkers, B = a set with 5% of utterances from all talkers, C = a set with 10% of utterances from all talkers.

F2B (6996 pitch accent tokens). The F2B data are listed as "development test" data in Table 2 because they were used to determine the best setting of a median-smoothing window (see text surrounding Fig. 2). The pitch accent and phrase boundary detectors described in Section 8 were trained and tested in a round-robin fashion, meaning that each event detector was trained and tested three times, and the resulting recognition accuracies were averaged. In each experiment, one of the talkers (F1A, M1B, M2B) was used for development testing, and the other two were used for training; the talker F2B was always part of the training set. In each of the three round-robin experiments, the number of pitch accent tokens in the training corpus was approximately 11,000. In these experiments, the test speaker is called a "development test" speaker in Table 2 because the elements of the acoustic observation vector are selected based on cross-validation.

The HMM-based prosody-dependent speech recognition systems described in Sections 6 and 7 were trained and tested in a multi-speaker speaker-dependent fashion. From the data of each talker in the Radio Speech Corpus, 85% of all utterances were selected for training, 5% were selected for development test, and 10% were selected for evaluation testing. The development test set was used to language model stream weights for the experiments described in Section 7. Section 6 did not use tunable stream weights, so the development test set was added to the training corpus. The decision to train and test in a speaker-dependent fashion rather than a speaker-independent fashion was based on an estimate of the training database size required to train a baseline prosody-independent HMM system with 48 phone models, 3 states/phone, 3 diagonal-covariance mixtures/state, and a 39-dimensional observation vector: according to one very rough estimate, a system with this description requires 180 min of training data. It would probably be possible to train and test the baseline speech recognizer in a round robin fashion, by including six talkers in the training set at all times, and by reserving the four medium-sized talkers (F1A, F3A, M1B, M2B) as test databases one after another, but we have not yet been able to develop training scripts to implement this testing method. All word error rate results reported in this paper are therefore multi-speaker speaker-dependent results (trained and tested using the same seven talkers) rather than speaker-independent results.

All HMM-based training and testing experiments described in this paper made use of the manually generated word and prosodic transcriptions and the automatically generated phone and part-of-speech transcriptions provided with the Radio Speech Corpus. Phone transcriptions were used to initialize phone models for each baseline prosody-independent HMM, and word transcriptions were used to re-estimate the prosody-independent HMM. Prosodic and part-of-speech transcriptions were used to train the prosody-dependent acoustic models, the prosody-dependent language models, and the joint prosodic-syntactic language models.

3. Methods: Overview

The experiments described in this paper are selected, from a large number of our recent published and unpublished experiments, in order to demonstrate the synergy of prosody and speech recognition. The sections of this paper are arranged with the intention of demonstrating that (a) the error rate of prosodic event detection drops when the word string is known, (b) the speech recognition word error rate drops when the prosodic events are known, and (c) if neither the word string nor the prosodic event tags are known a priori, then the error rates of both types of recognition may be reduced by using a simultaneous recognition paradigm that we call *prosody-dependent speech recognition*.

The reduction of word error rate is addressed in Sections 4 and 7. Section 4 gives a theoretical argument, demonstrating a condition that must hold in order for prosody-dependent recognition to reduce the WER of the recognizer. Section 7 provides experimental results demonstrating that prosody-dependent speech recognition can reduce WER on the Radio Speech Corpus by 12.5% relative. The reduction of prosodic event detection error is addressed in Sections 5–8. These sections discuss prosodic event detection under three conditions:

no knowledge of the word string (Sections 5 and 6), perfect a priori knowledge of the word string and phoneme alignment times (Section 8), and simultaneous recognition of prosody and the word string with no prior knowledge of either (Section 7).

All of the speech recognition systems described in this paper are conceptual variants of a core dynamic Bayesian network (DBN) model, shown in Fig. 1. Fig. 1 should be read as a conceptual summary of all of the speech recognition systems to be described in Sections 5–8, and not as a literal implementation guide. In most cases (specifically, in Sections 6–8), the actual implementation was designed by compiling the desired sections of Fig. 1 to an explicit-duration HMM (EDHMM), and by linking the EDHMM to neural networks trained to estimate acoustic or language model probabilities. Further implementation details are given in the sections that follow.

The key hidden variables in Fig. 1 are the prosody-dependent word, the prosody-dependent allophone, and the acoustic observation streams. The prosody-dependent word string may be further decomposed into an orthographic word string $W = [w_1, \ldots, w_M]$, a prosodic tag string $P = [p_1, \ldots, p_M]$, and a syntactic tag string $S = [s_1, \ldots, s_M]$. The goal of the systems described in this paper is to choose an optimal word string $\widehat{W}$ or an optimal prosodic tag string $\widehat{P}$ by maximizing the modeled posterior probability:

$$\widehat{W} = \arg\max_{W} \max_{S,P} p(W, S, P, O) \tag{4}$$

$$\widehat{P} = \arg\max_{P} \max_{W,S} p(W, S, P, O) \tag{5}$$

where $O = [\vec{o}_1, \ldots, \vec{o}_T]$ is the set of all acoustic observations. As in most large vocabulary speech recognition systems, Eq. (5) is implemented by way of an intermediate sequence of phoneme

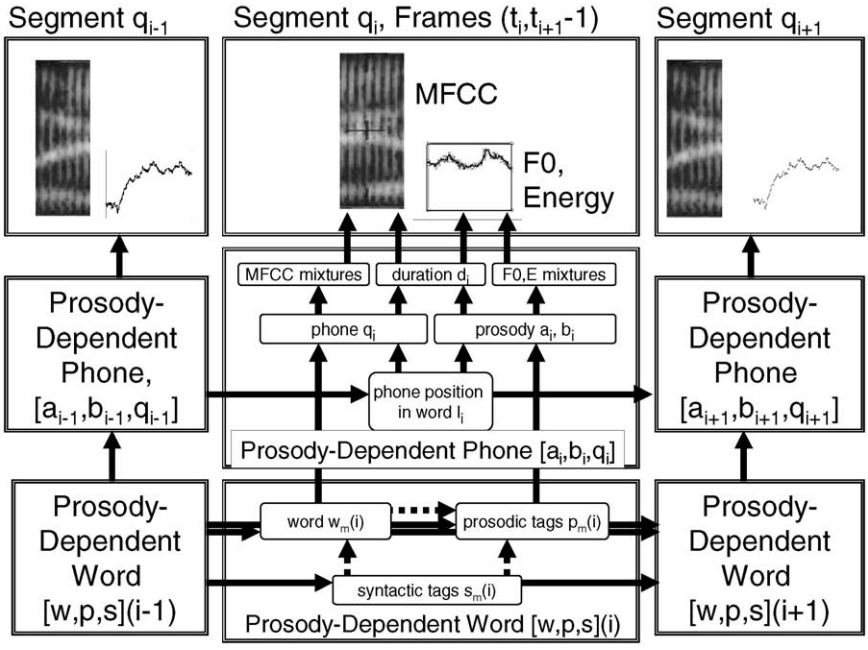

Fig. 1. A dynamic Bayesian network model of the interaction among prosody, words, syntax, phoneme durations, pitch, and the mel-frequency cepstrum. Columns in the figure represent consecutive phoneme segments (q_{i-1}, q_i, q_{i+1}). The super-variable "prosody-dependent phone" can be factored into the phone position in word (l_i), the ARPABET phone label (q_i), the phoneme-level prosodic tags (a_i, b_i), the phoneme duration (d_i), and the MFCC and F0 mixture indices. The super-variable "prosody-dependent word" can be factored into an orthographic word label (w_m), a set of prosodic tags (p_m), and a set of syntactic tags (s_m).

labels, $Q = [q_1, \ldots, q_L]$, and phoneme-level prosody tags, $[A, B] = [a_1, b_1, \ldots, a_L, b_L]$, thus

$$p(W, S, P, O) \approx \max_{Q} p(O|Q, A, B)p(Q, A, B|W, P)p(W, P|S)p(S) \quad (6)$$

Stochastic dependencies among variables are indicated by arrows in the graph. Thus, for example, there are three dashed arrows connecting the variables "syntactic tag", "prosodic tag", and "word"; these arrows are dashed because the dependencies that they encode are present in some of the implemented systems, and absent in others. In formal notation, the lack of an arrow between "prosodic tag" and "word" indicates that these two variables are conditionally independent of one another, given knowledge of the syntactic tags. If the mth word, prosodic tag, and syntactic tag are denoted as (w_m, p_m, s_m), then the absence of an arrow between word and prosodic tag implies the following equation:

$$p(w_m, p_m|s_m, w_{m-1}, p_{m-1}) = p(w_m|s_m, w_{m-1})p(p_m|s_m, p_{m-1}) \quad (7)$$

Experiments described in Section 7 will show that the model given in Eq. (7) yields lower WER than the tested alternatives.

Fig. 1 represents a dynamic Bayesian network with variable length outputs. The number of centisecond frames generated by each phone state is an explicit random variable, dependent on the phone identity, and dependent on the prosodic variable "lengthening". The variable "lengthening" encodes the distinction among different prosodic, syllabic, and word boundary contexts that may cause lengthening or shortening of phonemes. Explicit-duration hidden Markov models were implemented using Ferguson's algorithm (Ferguson, 1980); the HTK extensions are posted at http://www.ifp.uiuc.edu/speech/software/.

4. Synergy of acoustic and syntactic models of prosody

Katagiri et al. (1998) have shown that the expected sentence error rate (SER) of a speech recognizer can be written as a non-linear function of the log likelihood difference between the true word sequence, W_T, and all competing false word sequences $\widehat{W}_i$. The expected value of SER is given as:

$$E[SER] = E_{W_T,O}\left\{u(\log \max_i \eta_i)\right\} \quad (8)$$

where $u(x)$ is the unit step function, and η_i is

$$\eta_i = \frac{p(O, \widehat{W}_i)}{p(O, W_T)} = \frac{p(O|\widehat{W}_i)}{p(O|W_T)} \times \frac{p(\widehat{W}_i)}{p(W_T)} \quad (9)$$

The expected value of word error rate (WER) can be written in a form similar to Eq. (8), but only if several intermediate terms and functions are first defined. Let $W_T = [w_{T1}, \ldots, w_{TM}]$ be the correct transcription, while $\widehat{W}_i = [\widehat{w}_{i1}, \ldots, \widehat{w}_{iM}]$ is the ith incorrect transcription (among all possible incorrect transcriptions). In order to account for insertions and deletions, either w_{Tm} or $\widehat{w}_{im}$ may be NULL; the number of non-NULL words in the correct transcription is $M_T \leqslant M$. Let O_{Tm} be the sequence of observation vectors aligned with w_{Tm} in the maximum-likelihood alignment of observation sequence O with correct transcription W_T; likewise, O_{im} is the observation sequence aligned with $\widehat{w}_{im}$. Finally, let $W_{T,1:m} = [w_{T1}, \ldots, w_{Tm}]$ and $\widehat{W}_{i,1:m} = [\widehat{w}_{i1}, \ldots, \widehat{w}_{im}]$ be the m-word initial subsequences of W_T and $\widehat{W}_i$. Given these definitions, expected WER is

$$E[\mathrm{WER}] = E_{W_T,O}\left\{\frac{1}{M_T}\sum_{m=1}^{M} u\left(\log\ \mathrm{maxgeom}_{i,m}\eta_{im}\right)\right\} \quad (10)$$

where

$$\eta_{im} = \frac{p(O_{im}|\widehat{w}_{im})}{p(O_{Tm}|w_{Tm})} \times \frac{p(\widehat{w}_{im}|\widehat{W}_{i,1:(m-1)})}{p(w_{Tm}|W_{T,1:(m-1)})} \quad (11)$$

and where the non-linear function $\mathrm{maxgeom}_{i,m}\eta_{im}$ selects the variable η_{im} whose sequence $[\eta_{i1}, \ldots, \eta_{iM}]$ has maximum geometric average.

The expected WER of a prosody dependent speech recognizer can be written in a fashion similar to Eq. (10), resulting in

$$E[\mathrm{WER};P] = E_{W_T,O}\left\{\frac{1}{M_T}\sum_{m=1}^{M} u\left(\log \mathrm{maxgeom}_{i,m}\eta'_{im}\right)\right\} \quad (12)$$

where

$$\eta'_{im} = \frac{\sum_{\widehat{P}} p(O_{im}, \widehat{w}_{im}, \widehat{P} | \widehat{W}_{i,1:(m-1)})}{\sum_{\widehat{P}} p(O_{Tm}, w_{Tm}, \widehat{P} | W_{T,1:(m-1)})} \approx \frac{p(O_{im}|\widehat{w}_{im}, \hat{p}_{im})}{p(O_{Tm}|w_{Tm}, p_{Tm})} \times \frac{p(\widehat{w}_{im}, \hat{p}_{im}|\widehat{W}_{i,1:(m-1)}, \widehat{P}_{i,1:(m-1)})}{p(w_{Tm}, p_{Tm}|W_{T,1:(m-1)}, P_{T,1:(m-1)})}, \quad (13)$$

$P_T = [p_{T1}, \ldots, p_{TM}]$ is the sequence of word-synchronized prosodic tags that maximize $p(O, W_T, \widehat{P})$, and $\widehat{P}_i = [\hat{p}_{i1}, \ldots, \hat{p}_{iM}]$ is the prosodic tag sequence that maximizes $p(O, \widehat{W}_i, \widehat{P})$.

One of the goals of this paper is to design a prosody-dependent speech recognizer with a lower WER than the prosody-independent baseline system. From Eqs. (10) and (12), the objective of lower expected WER is satisfied if and only if

$$0 < E[\text{WER}] - E[\text{WER}; P] = E_{W_T,O}\left\{\frac{1}{M_T}\sum_{m=1}^{M}\left(u\left(\log \text{maxgeom}_{i,m}\eta_{im}\right) - u\left(\log \text{maxgeom}_{i,m}\eta'_{im}\right)\right)\right\} \quad (14)$$

Katagiri et al. (1998) suggested, for analytic purposes, replacing the step function in Eqs. (8)–(14) with a differentiable function such as a sigmoid or an identity. Using the sigmoid approximation ($u(x) \approx 1/(1 + e^{-x})$), Eq. (14) can be simplified to

$$E_{W_T,O}\left\{\frac{1}{M_T}\sum_{m=1}^{M}\text{maxgeom}_{i,m}\eta'_{im}\right\} < E_{W_T,O}\left\{\frac{1}{M_T}\sum_{m=1}^{M}\text{maxgeom}_{i,m}\eta_{im}\right\} \quad (15)$$

Eq. (15) demonstrates that the sigmoid approximation of the expected WER of a prosody-dependent speech recognizer is guaranteed to be less than that of the baseline system if $\eta_{im} > \eta'_{im}$ most of the time, where the phrase "most of the time" is quantified by the arithmetic average over m of the non-linear, non-local maxgeom function over i. We have previously shown (Chen et al. (in press); Chen and Hasegawa-Johnson (2004)) that mutual information between the word sequence and the acoustic signal is increased, by an explicit model of prosody, under similar circumstances; it is also possible to show that similar requirements lead to a decrease in sentence error rate. In all three cases, error rate is decreased, and mutual information is increased, if $\eta_{im} > \eta'_{im}$ most of the time, where the difference among these three criteria is in the precise definition of the phrase "most of the time". Re-arranging terms, the condition $\eta_{im}/\eta'_{im} > 1$ may be written

$$\left(\frac{p(p_{Tm}|W_{T,1:m}, P_{T,1:(m-1)})}{p(\hat{p}_{im}|\widehat{W}_{i,1:m}, \widehat{P}_{i,1:(m-1)})}\right) \times \left(\frac{p(O_{Tm}|w_{Tm}, p_{Tm})/p(O_{Tm}|w_{Tm})}{p(O_{im}|\widehat{w}_{im}, \hat{p}_{im})/p(O_{im}|\widehat{w}_{im})}\right) > 1 \quad (16)$$

Eq. (16) expresses the fraction η_{im}/η'_{im} as the product of two terms.

The first term on the left in Eq. (16) expresses the improvement, due to prosody, in the selectivity of the language model. It is greater than one, for example, when the true word sequence is uttered with a highly predictable prosodic pattern, thus $p(p_{Tm}|W_{T,1:m}, P_{T,1:(m-1)}) > p(\hat{p}_{im}|\widehat{W}_{i,1:m}, \widehat{P}_{i,1:(m-1)})$. 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 (Kompe (1997); Wightman and Ostendorf (1994)) has shown that intonational phrase boundaries are well predicted by N-gram word sequence or part-of-speech sequence statistics.

The second term on the left expresses the improvement, due to prosody, in the selectivity of the acoustic model. It is greater than one, for example, when knowledge of the prosodic tag p_{Tm} affiliated with word w_{Tm} increases the modeled likelihood of observation sequence O_{Tm}, but none of the possible prosodic tags associated with word sequence $\widehat{w}_{im}$ provide any similar acoustic modeling benefit. This term may be maximized by selectively modeling only those acoustic features whose distributions are well predicted by prosodic labeling. Beckman (1996) suggests that speaker-normalized fundamental frequency (f_0) is well predicted by the location of pitch accents, while Wightman et al. (1992) suggest that normalized phoneme duration is well predicted by the location

of intonational phrase boundaries. Cole et al. (2003) describe the prosody-dependent modification of the acoustic–phonetic features 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 Eq. (15) may therefore be explained in the following words: E[WER; P] < E[WER] in their sigmoid approximations 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.

5. Text-independent pitch accent recognition

Is the acoustic signal sufficient to determine whether or not a syllable is accented? Specifically, can a machine learning algorithm correctly transcribe the locations of pitch accents without prior information about word string, part of speech, or word boundary times? This section considers the problem of pitch accent recognition as a special case of the general problem of context-dependent, non-parametric dynamic contour recognition: given raw acoustic information as a function of time (specifically, F0, energy, spectral tilt, and zero-crossing rate), the task of the recognizer is to correctly mark the beginning and end time of every pitch-accented syllable, with no explicit internal model of words or syntax. The method proposed here consists of two steps: (1) F0 and energy contour normalization, and (2) pitch accent recognition using a time-delay neural network (TDNN; Waibel et al., 1989) or time-delay recursive neural network (TDRNN; Kim, 1998).

For the experiments presented in this section, F0 and energy contour normalization consists of the following steps. First, the fundamental frequency $F_0(t)$ and probability of voicing are extracted using the `formant` program in Entropic XWAVES. Second, we eliminated pitch doubling and halving errors by eliminating $F0$ that falls into the doubling and halving clusters of a three mixture Gaussian model whose mixture component means are restricted to $1/2\mu$, μ, and 2μ, where μ is the estimated utterance mean F_0 (Sönmez et al., 1998). We then normalize F_0 by μ and convert it to log scale:

$$\widehat{F}_0(t) = \max\left(0.2, \log\left(\frac{F_0(t)}{\mu} + 1\right)\right). \tag{17}$$

Eq. (17) is intended to mimic Fujisaki's $\log(f_0/\min f_0)$ parameterization (Fujisaki and Hirose, 1984; Hirai et al., 1997); in our experiments we found that estimates of the mean pitch are less sensitive to pitch tracking errors than estimates of the $\min f_0$, thus we find that Eq. (17) is less sensitive to pitch tracking errors than Fujisaki's parameterization. To eliminate unreliable $\widehat{F}_0$ measures, those with probabilities of voicing smaller than a heuristic threshold are replaced by the linear interpolated values $\widetilde{F}_0$ based on the $\widehat{F}_0$ that have probabilities of voicing greater than the threshold. Similarly, energy is normalized using

$$\widetilde{E}_0(t) = \max\left(-3, \log\left(\frac{E0(t)}{\eta}\right)\right), \tag{18}$$

where η is the utterance maximum energy. The clipping thresholds (0.2 for $\widehat{F}_0(t)$, -3 for $\widetilde{E}_0(t)$) were chosen experimentally for optimum results.

Pitch accent recognition is implemented using hidden Markov models (HMM), time-delay neural networks (TDNN), and time-delay recursive neural networks (TDRNN). All recognition architectures view a two-dimensional observation once per 10 ms frame, consisting of $[\widetilde{F}_0(t), \widetilde{E}_0(t)]$, where t is the frame index. The TDNN is configured with 32 input units, observing 2 input streams at each of 16 frames, specifically $[\widetilde{F}_0(t-15), \widetilde{E}_0(t-15), \ldots, \widetilde{F}_0(t), \widetilde{E}_0(t)]$. The TDRNN is configured with only 28 input units, observing $[\widetilde{F}_0(t-13), \widetilde{E}_0(t-13), \ldots, \widetilde{F}_0(t), \widetilde{E}_0(t)]$. In addition to time delays at the input, the TDRNN also has a recursive long-term

context node with 2 streams × 18 time delays = 36 units. The TDNN uses one hidden layer with 20 hidden units; the TDRNN uses two hidden layers, the first containing 10 units, the second containing only 2 units. The TDRNN's recursive long-term context nodes are direct copies (without modification) of previous activation levels in the second hidden layer of the network; outputs of the long-term context node are observed by the first hidden layer. The TDNN has a total of 742 trainable parameters; the TDRNN has a total of 696 trainable parameters. Both neural networks were trained to imitate a "target function" that was set to 1.0 during the sonorant portion of every syllable transcribed with a pitch accent, and was set to 0.0 during all other frames. The networks were trained using all of the transcribed speech data (67 paragraphs containing 2078 pitch accents) from one female speaker (F1A), and tested with all of the transcribed speech data (164 paragraphs containing 6996 pitch accents) from another female speaker (F2B). About 90% of all accented syllables are aligned with a high pitch (H* or !H*), about 5% are aligned with a low pitch (L*), and about 5% were considered questionable by the transcribers (?*).

As a baseline, the neural network-based pitch accent recognizers were compared to an HMM-based recognizer, built using five three-state HMMs. These five HMMs each model a subset of the nine accent types used in the Radio Speech Corpus, as follows: the H* model represents accent types (H*, L + H*), the !H* model represents accent types (!H*, L + !H*, H + !H*), the L* model represents accent types (L*, L* + H), the ?* model represents ?* accents, and the UA model represents unaccented syllables (the test data, from speaker F2B, included a number of pitch accents whose type was not marked, and which were therefore labeled with a bare "*"; the training data from speaker F1A contained no such labels). The observation PDF in each state is a 10-component diagonal-covariance mixture Gaussian PDF with a six-dimensional feature vector comprising $\widetilde{F}_0(t)$, $\widetilde{E}_0(t)$, and their deltas and delta-deltas. The HMMs have 393 trainable parameters each, for a total of 2358 trainable acoustic parameters. Pitch recognition for the HMM (but not the MLP or TDRNN) is also constrained by a bigram pitch-event language model, for a total of $2358 + 5 \times 4 = 2398$ trainable parameters.

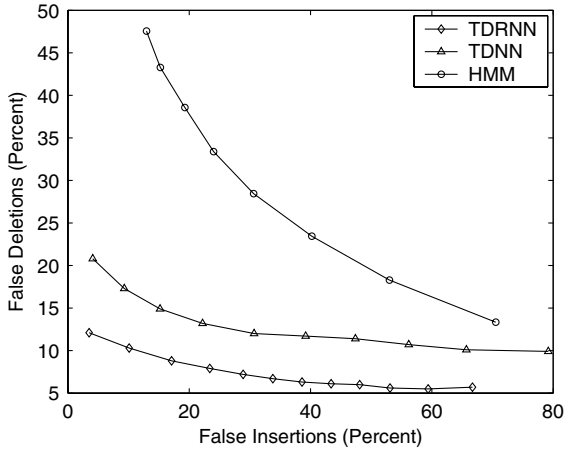

Fig. 2. Pitch accent detection errors: deletions vs. insertions, for three pitch accent recognizers operating with no explicit model of the word sequence. TDRNN = time-delay recursive neural network, TDNN = time-delay neural network, HMM = five-class mixture-Gaussian hidden Markov model.

All three recognizers (TDNN, TDRNN, and HMM) were trained using data from one female talker in the Radio Speech Corpus (F1A), and tested using data from a different female talker (F2B), thus all three systems were tested as speaker-independent, gender-dependent pitch accent detectors. All three systems produced an estimate of the probability of pitch accent in every frame; the raw probabilities produced by each classifier were then median smoothed in order to compute the final classifier output. By adjusting the length of the median smoothing window, it is possible to tune the classifier in order to trade off pitch accent insertions versus deletions. Fig. 2 shows deletion rate as a function of insertion rate for all fully trained recognizers. Equal error rate of the TDRNN is 10.2%, corresponding to less than 20% of words incorrectly labeled.

6. Text-independent intonational phrase boundary recognition

Intonational phrase boundaries are signalled by at least three types of cues: increased duration of

phonemes in the rhyme of the phrase-final syllable (Wightman et al. (1992)), a characteristic F0 movement called a boundary tone (Beckman and Elam (1994)), and increased glottalization of the phrase-initial phonemes (Dilley et al. (1996)). This section considers, in particular, the use of increased phoneme duration as a cue for the detection of intonational phrase boundaries. If phoneme boundaries and phoneme labels are available, it is possible to use duration cues directly; methods that require pre-existing segment boundaries have been considered by Wightman and Ostendorf (1994) and by Ostendorf and Ross (1997). If phoneme boundaries and phoneme labels are not available, it is necessary to simultaneously recognize the phoneme string and the relative lengthening of pre-boundary phonemes.

In a left-to-right hidden Markov model, the dwell time d_i of state i is an implicit random variable with a geometric PMF, i.e.,

$$p(d_i) = (1 - a_{ii}) a_{ii}^{d_i - 1} \tag{19}$$

where a_{ii} is the state's self-loop probability. If a hidden Markov model is composed of N consecutive states, all of which have the same self-loop probability a_{11} and with no skipping of states, then the total dwell time $D = d_1 + \cdots + d_N$ of the N-state model is a random variable with a gamma PDF, given by

$$p(D|N) = \frac{(D-1)!}{(N-1)!(D-N)!} a_{11}^{D-N} (1 - a_{11})^N \tag{20}$$

Crystal and House (1988) demonstrated that, if N is chosen to fit the distribution of phoneme durations observed in a large corpus, Eq. (20) can be an arbitrarily accurate model of true phoneme durations. In practical speech recognition systems, however, N is usually fixed at a small number such as $N = 3$.

Ferguson (1980) demonstrated an efficient training algorithm for an explicit-duration hidden Markov model (EDHMM), also known as a semi-Markov model. If $S = [s_1, \ldots, s_N]$ are the EDHMM state variables associated with phoneme q, and assuming that states in the model may not be skipped, then Ferguson's algorithm efficiently computes the following probability:

$$p(\vec{o}_1, \ldots, \vec{o}_T | q) = \prod_{i=1}^{N} p(d_i = \hat{t}_i - \bar{t}_i | s_i) \prod_{t=\bar{t}_i}^{\hat{t}_i} p(\vec{o}_t | s_i) \tag{21}$$

where $\bar{t}_i$ and $\hat{t}_i$ are the begin time and end time of state s_i, and $\vec{o}_t$ are the associated observation vectors. In previous work Chen et al. (in press), we demonstrated equations extending Ferguson's method to the Viterbi algorithm, implemented an EDHMM token-passing algorithm as an extension to HTK (the hidden Markov modeling toolkit; Young et al. (2002)), and demonstrated that these extensions may be used to model phrase-final lengthening at the end of intonational phrases.

In order to model phrase-final lengthening of individual phonemes, it is necessary to condition the state durations d_i on a prosodic context variable. In experiments reported here, the prosodic context variable may take on two different values: $b \in \{\text{phrase-final}, \text{non-final}\}$. In order to limit the complexity of the recognizer, we assume that only duration depends on b; given duration, the cepstral and other observation vectors are independent of b. The resulting observation PDF is given by

$$p(\vec{o}_1, \ldots, \vec{o}_T | q, b) = \prod_{i=1}^{N} p(d_i = \hat{t}_i - \bar{t}_i | s_i, b) \prod_{t=\bar{t}_i}^{\hat{t}_i} p(\vec{o}_t | s_i) \tag{22}$$

In a prosody-dependent EDHMM, the dependence of duration on prosody is implemented explicitly, exactly as shown in Eq. (22). In a prosody-dependent HMM, the dependence of duration on prosody is implemented implicitly, by giving each state a self-loop probability a_{ii} that depends on prosodic context, while the output vector probabilities $p(\vec{o}_t|s_i)$ are independent of prosody.

Can intonational phrase boundaries be recognized on the basis of phoneme durations alone, with no information about word content? Table 3 gives results using both an HMM and an EDHMM; the results suggest that intonational phrase boundaries cannot be accurately recognized based purely on phoneme-level acoustic information. The baseline models, in both cases, are three-state left-to-right hidden Markov models with three Gaussian mixtures per state. The

Table 3
Recognition error rate of intonational-phrase boundary-dependent allophones (top row) and of monophones (bottom row) in the Radio Speech Corpus

	HMM	EDHMM
Prosody-dependent allophone error rate (%)	75.4	74.6
Monophone error rate	49.0	48.1

models represent 96 prosody-dependent allophones: each of the 48 SPHINX monophones may occur in either intonational-phrase-final context or in phrase-non-final context. The phrase-final and non-final allophones of each phoneme share the same MFCC distributions $p(\vec{o}_t|s_i)$; only the duration distributions $p(d_i|s_i,b)$ depend on prosodic context. Models are trained using an arbitrarily selected 90% of the speech data in the Radio Speech Corpus, and are tested on the remaining 10%; training and test data contain the same talkers (recall that there are only seven talkers in the Radio Speech Corpus). Table 3 shows the results of four experiments. Monophone recognition accuracy is 51.0% using an HMM, and 51.9% using an EDHMM; these results are somewhat lower than the 73.7% monophone recognition accuracy we achieved using a similar HMM system on the TIMIT corpus (Omar and Hasegawa-Johnson, 2003), suggesting that the Radio Speech style is more difficult to recognize than the read speech in the TIMIT corpus. When we try to simultaneously correctly label both the monophone and the intonational phrase boundary position, accuracy drops to 24.6% with the HMM, and 25.4% using the EDHMM. Error analysis of these results suggests that many errors involve errors of both intonational phrase context and phoneme label: for example, a phrase-final short vowel, such as /IH/, may be mistaken for a corresponding non-final long vowel, such as /IY/. From these results, it appears that automatic recognition of intonational phrase boundaries requires some sort of constraint on the possible ordering of phonemes. The next two sections will consider two such constraints: first, the simultaneous recognition of words and prosody, and second, recognition of phrase boundaries given prior knowledge of word and phoneme labels and alignment times.

7. Simultaneous recognition of words and prosody

The relationship among syntax, prosody, and the word string is modeled in our system by a tagged language model. A tagged language model is an estimate of the probability $p(w_m,p_m,s_m|\text{history})$ where w_m is the mth word in the sentence, and p_m and s_m are its prosodic and syntactic tags, respectively. The amount of prosodically labeled data in the English language is not nearly sufficient to create a reliable maximum likelihood estimate of $p(w_m,p_m,s_m|\text{history})$, therefore we have experimented with three methods for estimating the language model probability: a backed-off prosodically-labeled bigram (with no encoding of syntax), and two factored language models.

A prosody-dependent bigram is an estimate of $p(w_m,p_m|w_{m-1},p_{m-1})$. 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). A prosodically tagged word may be encoded in 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). 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-g and 0-g (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.

An empirically superior estimate of the prosody-dependent bigram probability may be trained by explicitly modeling the relationship between the prosodic tag, p_k, and the syntactic tag, s_k Chen and

Hasegawa-Johnson (2003). The syntactic tag s_k specifies the part of speech of word w_k, and during second-pass decoding (given a complete sentence hypothesis), may also specify the position of word w_k relative to syntactic phrase and clause boundaries. By explicitly modeling syntactic tags, the prosody-dependent bigram probability may be written as

$$p(w_j,p_j|w_i,p_i)=\sum_{s_j,s_i}p(w_j,p_j,s_j,s_i|w_i,p_i) \tag{23}$$

$p(w_j,p_j,s_j,s_i|w_i,p_i)$ is proportional to the bigram probability of a syntactically and prosodically tagged vocabulary. This tagged bigram probability may be computed as

$$\begin{aligned}&p(w_j,p_j,s_j,s_i|w_i,p_i)\\&\quad\approx p(p_j|s_j,s_i,p_i)p(s_j,s_i|w_j,w_i)p(w_j|w_i,p_i)\end{aligned} \tag{24}$$

The approximation in Eq. (24) is valid if we assume that, first, prosody is independent of the word string given knowledge of syntax (reasonable because neither side of the equation has any explicit representation of dialog context), and second, that the syntactic tags are independent of prosody given knowledge of the word string (reasonable except for those cases when prosody may be used to resolve syntactic ambiguity, e.g. Price et al. (1991)). Under these assumptions, the tagged bigram probability factors into three terms. The first term, $p(p_j|s_j,s_i,p_i)$, may be robustly estimated from a relatively small corpus, because the syntactic tagset and the prosodic tagset are both much smaller than the vocabulary. The second term, $p(s_j,s_i|w_j,w_i)$, is the probability that a word sequence (w_i,w_j) implements syntactic tag sequence (s_i,s_j). Computation of this probability is simplified by appropriate choice of the syntactic tagset. During first-pass recognition, the syntactic tag s_i encodes only the part of speech of word w_i. In most cases, the word sequence (w_i,w_j) uniquely determines the POS sequence (s_i,s_j); the few common exceptions can be robustly estimated from a large text database with manual or automatic POS tags. During second-pass recognition, in an N-best rescoring paradigm, it is possible to assume that the recognizer is computing the prosody-dependent and syntax-dependent probability of a complete sentence transcription, $W=[w_1,\ldots,w_M]$. Given a complete transcription, it is possible to compute the maximum likelihood phrase-level parse of the sentence using a context-free grammar, and to augment the syntactic tag s_i with information about the position of the word in its surrounding phrase and clause. Like POS, this new syntactic information may be treated, by the prosody-dependent language model, as information uniquely determined by the hypothesized word sequence (w_i,w_j).

The third term in Eq. (24), $p(w_j|w_i,p_i)$, is a prosody-dependent semi-bigram probability. We have tested two variants of Eq. (24): one in which the probability $p(w_j|w_i,p_i)$ is estimated directly from the Radio Speech Corpus, using backed-off ML estimation, and one in which the probability is estimated using the following approximation:

$$\begin{aligned}p(w_j|w_i,p_i)&=\frac{p(p_i|w_j,w_i)p(w_j|w_i)}{p(p_i|w_i)}\\&\approx\frac{\sum_{s_i,s_j}p(p_i|s_i,s_j)p(s_i,s_j|w_i,w_j)p(w_j|w_i)}{\sum_{w_j}\sum_{s_i,s_j}p(p_i|s_i,s_j)p(s_i,s_j|w_i,w_j)p(w_j|w_i)}\end{aligned} \tag{25}$$

Table 4 describes performance of six different recognizers, each based on 48 monophone HMMs, each composed of an MFCC observation stream (three-mixture Gaussian) and a pitch observation stream (Gaussian), with explicit representation of duration probability density. Each row was created by training the named recognizer on about 90% of the ToBI-transcribed data in the Radio Speech Corpus (six talkers), and testing on the

Table 4
Word error rate (WER), accent error rate (AER), and intonational phrase boundary error rate (BER, in percent) with six different combinations of acoustic model (AM) and language model (LM)

AM	LM	WER	AER	BER
PI	PI	24.8	44.6	15.6
PD	PI	24.0	45.9	15.0
PI	PD bigram	24.3	23.1	14.5
PD	PD bigram	23.4	20.3	14.3
PD	PD semi-factored	21.7	20.3	14.2
PD	PD factored	22.9	19.7	13.4

PI = prosody independent (baseline), PD = prosody dependent. Accent and boundary error rates of the system with no prosody dependence are at chance.

remaining 10% (from the same six talkers). During testing, each recognizer output its best estimate of the complete lexical and prosodic transcription of the utterance. Word error rate was computed by comparing the lexical transcription to a reference using the program HResults, without considering the prosodic transcription; accent and boundary recognition error rates were computed by ignoring the lexical transcription. The system in the first row has no explicit representation of prosody. Accent and boundary recognition error rates of the first system are at chance for this database: 45% of words in this database are unaccented (55% are accented), and 16% are phrase-final. In the second system, the F0 stream is accent-dependent and the duration PMF is phrase-position dependent; all systems in this table use a prosody-independent MFCC stream. The third system uses a prosody-dependent bigram language model with no model of the acoustic correlates of prosody. The fourth system uses a prosody-dependent bigram, plus explicit models of accent-dependent pitch variation and phrase-final lengthening. The fifth system uses a semi-factored language model, meaning that $p(w_j,p_j|w_i,p_i)$ is factored, but $p(w_j|w_i,p_i)$ is not (Eq. (24)) is used, but not Eq. (25). The last system uses both Eqs. (24) and (25).

The results of Table 4 indicate that word error rate is only significantly improved if a prosody-dependent acoustic model and a prosody-dependent language model are combined. Prosody-dependent language modeling, alone, is sufficient for better-than-chance recognition of accents and boundaries; a prosody-dependent acoustic model, alone, is insufficient for any type of gain.

The last two rows of Table 4 present results obtained using the semi-factored and factored bigram language models. The word perplexities of the bigram, semi-factored, and factored language models, using the same test corpus as in Table 4, are 60, 54, and 47, respectively. The semi-factored model has significantly lower WER than the baseline bigram (21.7% vs. 23.4%), but not significantly lower boundary error rate (14.2% vs. 14.3%) or accent error rate (20.3% vs. 20.3%). The factored model has significantly improved boundary recognition error (13.4% vs. 14.3%), but not significantly improved WER (22.9% vs. 23.4%).

8. Prosody recognition given known word transcription

Consider the problem of recognizing the sequence of prosody tags, $P = [p_1, \ldots, p_M]$, given a sequence of word labels $W = [w_1, \ldots, w_M]$ with known alignment times. The optimal prosodic tag sequence is the sequence $\widehat{P}$ that maximizes the recognition probability:

$$\widehat{P} = \arg\max_P \prod_{m=1}^{M} p(y_m|w_m,p_m)p(p_m|\phi_m(W))^{\gamma}, \quad (26)$$

where y_m is the sequence of acoustic features that provide information about prosodic tag p_m (possibly including delta and delta-delta features), $\phi_m(W)$ is a function that describes all the information in W that may affect the prediction of p_m, and γ is the language model stream weight. Assuming the dependence of prosody on word strings is localized in a window of n words and is described by the syntactic roles of the words (primarily parts-of-speech) instead of the words themselves, then

$$\phi_m(W) = S_m = (s_{m-n+2}, s_{m-n+3}, \ldots, s_{m+n-1}) \quad (27)$$

where S_m is used to represent the set of syntactic information that affects the prediction of p_m. In the experiments in this section, s_m includes part-of-speech (ground truth for part-of-speech is provided by automatic transcriptions distributed with the Radio Speech Corpus), and information about syntactic phrase boundaries (syntactic phrase structure was estimated by applying Charniak's stochastic CFG parser Charniak (1994) to the orthographic transcriptions provided in the Radio Speech Corpus).

The probability $p(y_m|w_m,p_m)$ in Eq. (26) can be further expanded to the phoneme level:

$$p(y_m|w_m,p_m) = \sum_{Q_m,A_m,B_m} \left(\prod_{i=1}^{N_m} p(y_i|q_i,a_i,b_i)p(Q_m,A_m,B_m|w_m,p_m) \right) \quad (28)$$

where $p(Q_m,A_m,B_m|w_m,p_m)$ is a pronunciation model that computes the probability of a phoneme string $Q = [q_1, \ldots, q_{N_m}]$ and the accompanying

lengthening variables $B_m = [b_1, \ldots, b_{N_m}]$ and phoneme stress labels $A_m = [a_1, \ldots, a_{N_m}]$ given prosody dependent word token (w_m, p_m).

For simplicity, assume that the acoustic–prosodic features y_i are computed in the local context of each phoneme, possibly including information from a fixed window of frames on either side. Acoustic prosodic features include functions of the normalized pitch $\widetilde{F}_0(t)$, normalized energy $\widetilde{E}_0(t)$, and absolute phoneme duration. Absolute phoneme durations are computed using forced alignment of a prosody-independent hidden Markov model of the known word. $\widetilde{F}_0(t)$ and $\widetilde{E}_0(t)$ are computed using the methods described in Section 5. After normalization, a group of five features are computed as the base feature vector $\vec{x}_i$: (1) phoneme duration, (2) average phoneme duration over a window of three phonemes, (3) average $\widetilde{E}_0(t)$ over a window of three phonemes, (4) the delta of the three-phone-average of the phoneme-wise mean $\widetilde{F}_0(t)$, and (5) the delta of item number (4). There features are designed based on test and trial, and are found to give the best performance among a set of around 15 candidate features. After $\vec{x}_i$ is computed, the vector is rotated using principle component analysis (PCA) so that it can be better modeled by a diagonal covariance Gaussian PDF. The delta of the rotated feature vectors are attached to make up a 10-dimensional feature vector $\vec{y}_i$ for each phoneme segment. The acoustic observation PDF $p(y_i|q_i, a_i, b_i)$ is modeled using a mixture of diagonal-covariance Gaussians.

The language model $p(p_m|\phi_m(W))$ is implemented using a multilayer perceptron (MLP). The input nodes of the MLP observe a large number of binary and integer-valued variables encoding syntactic features of the five words $[w_{m-2}, \ldots, w_{m+2}]$, and including in particular the syntactic features identified as most useful in a previous study by Cohen (2004). The MLP has four output nodes, trained to estimate the four *a posteriori* probabilities $p(p_m = \text{unaccented}, \text{non-final}|\phi_m(W))$, $p(p_m = \text{accented}, \text{non-final}|\phi_m(W))$, $p(p_m = \text{unaccented}, \text{phrase-final}|\phi_m(W))$, $p(p_m = \text{accented}, \text{phrase-final}|\phi_m(W))$. The vector $\phi_m(W)$ encodes, for each word $w_i \in [w_{m-2}, \ldots, w_{m+2}]$, the part of speech (POS) of w_i, the number of syntactic phrases or clauses that end after word w_i, and the number of syntactic phrases or clauses that begin on word w_i. There are 33 POS tags, including "SIL" (silence), and the 32 Penn Treebank POS tags. The number of phrase and clause boundaries coincident with each word is computed by parsing the word string using Charniak's probabilistic context-free parser Charniak (1994). POS is encoded using 33 binary indicator variables per word, for a total of 165 binary indicator variables. The number of phrase boundaries ending and starting on each word are encoded in the form of two integer-valued features per word, for a total of 10 integer-valued features, thus the MLP has a total of 175 input nodes.

Data used in the experiments are extracted from 4 speakers in the Radio Speech Corpus: F1A, F2B, M1B and M2B. For each experiment, data from one speaker are used for test, and the other three are used to train models. The directory F2B is never left out because it contains the most training data. Results are given in Table 5.

As shown in Table 5, the acoustic model only (AM only) error rate (23.42%) is somewhat worse than the error rate achieved by Wightman and Ostendorf on a similar task (Wightman and Ostendorf, 1994), and is also slightly worse than the best results achieved using our TDRNN acoustic-only pitch accent detector (about 20%: Fig. 2); we have not yet combined the TDRNN with the pitch accent recognition system described here. All three sets of AM-only error rates (those in Table 5, those in Fig. 2, and those in Wightman and Ostendorf, 1994) are higher than the LM-only error rates cited in Table 5 (17.33%). The combination of both acoustic and language model information results in a pitch accent error rate comparable to the best

Table 5
Average accent, boundary and accent/boundary combined error rates (%) for acoustic model only (AM only), language model only (LM only) and acoustic model language model combined systems on the leave-one-speaker-out task on the Radio Speech Corpus

	Accent	Boundary	Acc. Bnd. combined
AM only	23.42	31.77	49.94
LM only	17.33	9.91	23.19
Combined	16.09	6.93	21.58

rates reported in previous studies of this corpus (16.09%, comparable to 16% reported in Wightman and Ostendorf, 1994).

Intonational phrase boundary recognition accuracy using only the acoustic model is considerably worse than chance (31.77%, compared to a chance level of 16%). Using only language model information, on the other hand, Table 5 achieves an intonational phrase boundary recognition error rate (9.91%) slightly but not significantly lower than that previously reported by Cohen (10.9%, Cohen, 2004), and considerably better than chance. The final intonational phrase boundary detection error rate (6.9%), achieved using both acoustic and language model information, appears to be the lowest error rate reported for this task, and closely approximates the inter-transcriber disagreement rate (reported to be about 5%, Ostendorf et al., 1995).

9. Discussion and conclusions

In this paper, a prosody dependent speech recognizer that models word and prosody in a unified probabilistic framework is proposed. A theoretical analysis is provided, showing that prosody dependent recognition can decrease the expected word error rate of a speech recognizer by utilizing the interaction between the acoustic model and language model. Four types of experiments are described. First, experiments are described that detect pitch accents with no knowledge of the words, phones, or phone alignment times in the utterance. The best such system, using a time-delay recursive neural network with fewer than 700 trainable parameters, achieves an equal error rate of close to 10% on a gender-dependent, speaker-independent accent recognition task. The second set of experiments attempts to recognize intonational phrase boundary position using knowledge of phonemes and phone alignment times, but with no information about the organization of phones into words; the resulting boundary error rates are worse than chance. The third set of experiments demonstrate that an explicit representation of prosody can reduce the word error rate of a multi-speaker speaker-dependent speech recognizer, but that statistically significant WER reductions depend on the simultaneous use of both a prosody-dependent acoustic model and a prosody-dependent language model. Additional improvements, in both perplexity and WER, can be obtained using a semi-factored language model, in which the relationship between prosody and the word sequence is at least partly mediated by syntactic tags. Finally, the fourth set of experiments uses complete knowledge of both the acoustics and the word string in order to derive the best possible prosodic transcription of an utterance. A speaker-independent intonational phrase boundary error rate of only 6.9% is achieved; this result is below the boundary error rates reported in other studies, and approximates quite closely the lowest reported inter-transcriber disagreement rates.

The experiments reported in this paper support the following two claims. First, accurate detection of pitch accents and intonational phrase boundaries requires information about word and phoneme alignment times, obtained either from prior knowledge (as in Section 8) or by simultaneously recognizing the words and prosody of the utterance (as in Section 7). Second, it is possible for an explicit model of prosody to reduce the word error rate of an automatic speech recognizer. Taken together, these results suggest that there is no such thing as a prosody-independent allophone in the Radio Speech Corpus. The spectral distribution of every allophone is adjusted to fit a particular prosodic context, therefore knowledge of the prosodic context aids recognition of the allophone, and conversely, knowledge of the allophone aids recognition of the prosody.

The immediate utility of the results presented in this paper is limited by the relatively small size of the speech corpus, and by the consequent use of a speaker-dependent rather than a speaker-independent speech recognizer. At present, there is no way to know whether these results can be generalized to larger corpora, to a speaker-independent speech recognizer, to speech styles other than radio speech, or to languages other than English. The methods used in this paper require training and testing on speech data with manual prosodic transcriptions. Manual prosodic transcriptions are expensive; to our knowledge, the 3.5-h Radio Speech Corpus is currently the largest

publicly available English-language speech database with manual prosodic transcriptions.

We are currently pursuing two methods for generalizing these experiments to other, larger speech corpora. First, we are experimenting with low-cost rapid prosodic transcriptions. Of all the distinctions labeled in the ToBI transcription standard, most speech recognition experiments in this paper made use of only two binary distinctions: intonational phrase-final vs. non-final, pitch accented vs. unaccented. Transcribers report being able to make these two distinctions very rapidly. Experiments are currently under way to determine whether rapid transcriptions of these two distinctions are accurate (high inter-transcriber agreement) and phonologically meaningful. If so, it may be possible to rapidly label larger speech corpora.

Second, we are experimenting with statistical and machine learning methods that may be able to generalize prosodic knowledge across speech corpora. In a factored acoustic model, for example, it is possible to train the different components of the acoustic model on different corpora. The pitch stream may be amenable to training using prosodically labeled data, while the MFCC stream is trained using a much larger (and perhaps more task-specific) training corpus. Allophone duration models may also be amenable to parameterization, e.g., by modeling the effect of phrase-final lengthening using a manner-class-dependent time-scaling operation (Klatt, 1976).

The experimental tests reported in this paper may be interpreted as an existence proof: under reasonably favorable conditions (manual transcriptions, speaker-dependent recognition), it is possible to use explicit models of prosodic allophony and of the interaction between prosody and syntax to reduce the word error rate of a speech recognizer, and to reduce the error rate of automatic prosodic transcription. The task of our ongoing research is to generalize this result.

References

Batliner, A., Kießling, A., Kompe, R., Niemann, H., Nöth, E., 1997. Can we tell apart intonation from prosody (if we look at accents and boundaries)? In: Proc. ESCA Intonation Workshop, Athens, pp. 39–42.

Beckman, M., 1996. The parsing of prosody. Language Cognitive Processes 11 (1), 17–67.

Beckman, M.E., Elam, G.A., 1994. Guidelines for ToBI labelling. Technical report, Ohio State University. Available from <http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html>.

Borys, S., 2003. Recognition of prosodic factors and detection of landmarks for automatic speech recognition, bachelor's thesis, University of Illinois at Urbana-Champaign.

Charniak, E., 1994. Statistical Language Learning. MIT Press, Cambridge, MA.

Chavarria, S., Yoon, T., Cole, J., Hasegawa-Johnson, M., 2004. Acoustic differentiation of ip and IP boundary levels: Comparison of L– and L–L% in the switchboard corpus. In Proc. SpeechProsody, Nara, Japan.

Chen, K., Hasegawa-Johnson, M., 2003. Improving the robustness of prosody dependent language modeling based on prosody syntax cross-correlation. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).

Chen, K., Hasegawa-Johnson, M., 2004. How prosody improves word recognition. In: Proc. SpeechProsody, Nara, Japan.

Chen, K., Borys, S., Hasegawa-Johnson, M., 2003a. Prosody dependent speech recognition with explicit duration modeling at intonational phrase boundaries. In: Proc. EUROSPEECH, Geneva, pp. 393–396.

Chen, K., Hasegawa-Johnson, M., Kim, S.-S., 2003b. An intonational phrase boundary and pitch accent dependent speech recognizer. In: Internat. Conf. on Syst., Cybernet., Intell. (SCI). Orlando, FL.

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

Chen, K., Hasegawa-Johnson, M., Cohen, A., Cole, J., 2004b. A maximum likelihood prosody recognizer. In: Proc. SpeechProsody, Nara, Japan.

Chen, K., Hasegawa-Johnson, M., Cohen, A., Borys, S., Kim, S.-S., Cole, J., Choi, J.-Y., in press. Prosody dependent speech recognition on radio news. IEEE Trans. Speech Audio Process.

Choi, H., Cole, J., Kim, H., 2003. Acoustic evidence for the effect of accent on CV coarticulation in radio news speech. In: Proc Texas Linguistics Conf. Univ Texas at Austin.

Cohen, A., 2004. A survey of machine learning methods for predicting prosody in radio speech. Master's thesis, University of Illinois at Urbana-Champaign.

Cole, J., Choi, H., Kim, H., Hasegawa-Johnson, M., 2003. The effect of accent on the acoustic cues to stop voicing in radio news speech. In: Internat. Conf. Phonet. Sci.

Crystal, T.H., House, A.S., 1988. Segmental druations in connected-speech signals: Current results. J. Acoust. Soc. Amer. 83, 1553–1573.

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

Dilley, L., Shattuck-Hufnagel, S., Ostendorf, M., 1996. Glottalization of word-initial vowels as a function of prosodic structure. J. Phonet. 24, 423–444.

Ferguson, J.D., 1980. Variable duration models for speech. In: Ferguson, J. (Ed.), Proc. Symp. Applic. Hidden Markov Models to Text and Speech. Princeton University Press, Princeton, NJ, pp. 143–179.

Ferrer, L., Shriberg, E., Stolcke, A., 2003. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In: Proc. ICASSP. pp. 608–611.

Fougeron, C., Keating, P.A., 1997. Articulatory strengthening at edges of prosodic domains. J. Acoust. Soc. Amer. 101 (6), 3728–3740.

Fujisaki, H., Hirose, K., 1984. Analysis of voice fundamental frequency contours for declarative sentence of Japanese. J. Acoust. Soc. Jpn. 5 (4), 233–242.

Godfrey, J., Holliman, E., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: Proc. ICASSP. pp. 517–520.

Greenberg, S., Hitchcock, L., May 2001. Stress-accent and vowel quality in the Switchboard corpus. In: NIST Large Vocabulary Continuous Speech Recognition Workshop, Linthicum Heights, MD.

Hahn, L., 1999. Native speakers' reactions to non-native stress in english discourse. Ph.D. Thesis, University of Illinois at Urbana-Champaign.

Hasegawa-Johnson, M., Cole, J., Shih, C., Chen, K., Cohen, A., Chavarria, S., Kim, H., Yoon, T., Borys, S., Choi, J.-Y., 2004. Speech recognition models of the interdependence among syntax, prosody, and segmental acoustics. In: HLT/NAACL Workshop on Linguist. Other Higher-Level Knowledge Speech Process.

Hirai, T., Iwahashi, N., Higuchi, N., Sagisaka, Y., 1997. Automatic extraction of f_0 control rules using statistical analysis. In: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), Progress in Speech Synthesis. Springer-Verlag, New York, pp. 333–346.

Hirschberg, J., Nakatani, C., 1998. Acoustic indicators of topic segmentation. In: Proc. Internat. Conf. on Spoken Language Process.

Hombert, J., 1978. Consonant types, vowel quality and tone. In: Fromkin, V. (Ed.), Tone: A Linguistic Survey. pp. 77–112.

Katagiri, S., Juang, B.-H., Lee, C.-H., 1998. Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method. Proc. IEEE 86 (11), 2345–2373.

Kent, Netsell, 1971. Effects of stress contrasts on certain articulatory parameters. Phonetica. 24, 23–44.

Kim, S.-S., 1998. Time-delay recurrent neural network for temporal correlations and prediction. Neurocomputing 20, 253–263.

Kim, H., Cole, J., Choi, H., Hasegawa-Johnson, M., 2004a. The effect of accent on acoustic cues to stop voicing and place of articulation in radio news speech. In: Proc. SpeechProsody, Nara, Japan.

Kim, S.-S., Hasegawa-Johnson, M., Chen, K., 2004b. Automatic recognition of pitch movements using multi-layer perceptron and time-delay recursive neural network. IEEE Signal Process. Lett. 11 (7), 645–648.

Klatt, D.H., 1976. Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. J. Acoust. Soc. Amer. 59 (5), 1208–1221.

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

Lee, K.-F., Hon, H.-W., 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoust., Speech, Signal Process. 37 (11), 1641–1648, November.

Liu, Y., Shriberg, E., Stolcke, A., 2003. Automatic disfluency identification in conversational speech using multiple knowledge sources. In: Proc. EUROSPEECH.

Omar, M.K., Hasegawa-Johnson, M., 2003. Approximately independent factors of speech using non-linear symplectic transformation. IEEE Trans. Speech and Audio Process. 11 (6), 660–671.

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

Ostendorf, M., Ross, K., 1997. A multi-level model for recognition of intonation labels. In: Computing Prosody: Computational Models for Processing Spontaneous Speech. Springer-Verlag, Inc., New York.

Ostendorf, M., Shafran, I., Shattuck-Hufnagel, S., Carmichael, L., Byrne, W., 2002. A prosodically labeled database of spontaneous speech. In: Proc. ISCA Tutorial Res. Workshop on Prosody in Speech Recognition Understand., Red Bank, NJ.

Pitrelli, J.F., Beckman, M., Hirschberg, J., 1994. Evaluation of prosodic transcription labeling reliability in the TOBI framework. In: Proc. Internat. Conf. Spoken Language Process.

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

Ren, Y., Kim, S.-S., Hasegawa-Johnson, M., Cole, J., 2004. Speaker-independent automatic detection of pitch accent. In: Proc. SpeechProsody, Nara, Japan.

Shriberg, E., Stolcke, A., 2004. Direct modeling of prosody: An overview of applications in automatic speech processing. In: Proc. SpeechProsody.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., 1992. TOBI: A standard for labeling English prosody. In: Proc. Internat. Conf. Spoken Language Process.

Sluijter, A.M.C., van Heuven, V.J., Pacilly, J.J.A., 1997. Spectral balance as a cue in the perception of linguistic stress. J. Acoust. Soc. Amer. 101, 503–513.

Sönmez, K., Shriberg, E., Heck, L., Weintraub, M., 1998. Modeling dynamic prosodic variation for speaker

verification. In: Proc. Internat. Conf. Spoken Language Process., pp. 3189–3192.

Stolcke, A., Shriberg, E., Hakkani-Tür, D., Tür, G., 1999. Modeling the prosody of hidden events for improved word recognition. In: Proc. EUROSPEECH, pp. 307–310.

Taylor, P., 2000. Analysis and synthesis of intonation using the Tilt model. J. Acoust. Soc. Amer. 107 (3), 1697–1714.

van Kuijk, D., Boves, L., 1999. Acoustic characteristics of lexical stress in continuous telephone speech. Speech Comm. 27, 95–111.

Vergyri, D., Stolcke, A., Gadde, V.R., Ferrer, L., Shriberg, E., 2003. Prosodic knowledge sources for automatic speech recognition. In: Proc. ICASSP.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J., 1989. Phoneme recognition using time-delay neural networks. Trans. Acoust. Speech Sig. Proc. 37, 328–339.

Wightman, C., Ostendorf, M., 1994. Automatic labeling of prosodic patterns. IEEE Trans. Speech Audio Process. 2 (4), 469–481, Oct.

Wightman, C., Shattuck-Hufnagel, S., Patti Price, M.O., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. J. Acoust. Soc. Amer. 91 (3), 1707–1717, March.

Yoon, T., Chavarria, S., Cole, J., Hasegawa-Johnson, M., 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using tobi. In: Proc. Internat. Conf. Spoken Language Process.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. The HTK Book. Cambridge University Engineering Department, Cambridge, UK.

Zue, V., Seneff, S., Glass, J., 1990. Speech database development at MIT: TIMIT and beyond. Speech Comm. 9, 351–356.