Unsupervised Adaptation of Categorical Prosody Models for Prosody Labeling and Speech Recognition

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan, *Senior Member, IEEE*

***Abstract*—Automatic speech recognition (ASR) systems rely almost exclusively on short-term segment-level features (MFCCs), while ignoring higher level suprasegmental cues that are characteristic of human speech. However, recent experiments have shown that categorical representations of prosody, such as those based on the Tones and Break Indices (ToBI) annotation standard, can be used to enhance speech recognizers. However, categorical prosody models are severely limited in scope and coverage due to the lack of large corpora annotated with the relevant prosodic symbols (such as pitch accent, word prominence, and boundary tone labels). In this paper, we first present an architecture for augmenting a standard ASR with symbolic prosody. We then discuss two novel, unsupervised adaptation techniques for improving, respectively, the quality of the linguistic and acoustic components of our categorical prosody models. Finally, we implement the augmented ASR by enriching ASR lattices with the adapted categorical prosody models. Our experiments show that the proposed unsupervised adaptation techniques significantly improve the quality of the prosody models; the adapted prosodic language and acoustic models reduce binary pitch accent (presence versus absence) classification error rate by 13.8% and 4.3%, respectively (relative to the seed models) on the Boston University Radio News Corpus, while the prosody-enriched ASR exhibits a 3.1% relative reduction in word error rate (WER) over the baseline system.**

***Index Terms*—Categorical prosody models, lattice enrichment, speech recognition, unsupervised adaptation.**

I. Introduction

PROSODY refers to specific patterns of rhythm, intonation, and lexical stress characteristic of human speech. These patterns are conveyed through modulation of the fundamental frequency (F0) contour, selective emphasis on certain words or syllables, and durational cues such as vowel lengthening, pauses, and hesitation. These acoustic correlates of prosody are referred to as *suprasegmental cues*, since they are usually associated with linguistic elements larger than phonemes, such as syllables, words, phrases, or even entire utterances; traditional speech processing and automatic speech recognition (ASR) systems typically operate at the segmental level and ignore such suprasegmental information. While the prosody of an utterance conveys information complementary to the segment-level spectral features used in many ASR systems, it is often difficult to exploit due to its suprasegmental nature. Moreover, the acoustic correlates of prosody exhibit high variability depending on a variety of factors, including context and the speaker's emotional state; the link between them and linguistic elements (typically words) is language specific, and for American English, tenuous at best. This makes it difficult to integrate prosody within spoken language systems, except in ad-hoc ways for very specific applications.

Categorical representations of prosody attempt to solve this problem by encoding prosodic events using a symbolic alphabet. Tones and Break Indices (ToBI) [1], INTSINT [2], and iViE [3] are all examples of symbolic prosody annotation standards well-known in the community. The advantages of adopting a symbolic representation of prosody over continuous-valued acoustic–prosodic features are twofold.

1) Symbolic representation of prosodic events greatly reduces the variability associated with acoustic–prosodic features and makes it easier for automatic learning algorithms to derive relationships between said events and other linguistic entities (e.g., words, syntactic boundaries).
2) Many of these annotation schemes (e.g., ToBI) are linguistically motivated; the prosodic transcription is thus correlated with linguistic elements (e.g., ToBI pitch accent is closely related to syllable stress and word prominence).

As a result, categorical representations of prosody are far more amenable to systematic integration within spoken language systems than the corresponding acoustic–prosodic features. Indeed, previous work on automatic detection and labeling of prosodic events in speech [4] suggests a strong co-occurence of lexical items (syllables, words, part-of-speech tags) and specific categorical prosodic events. For instance, content words are much more likely to be associated with pitch accent events than function words. Similarly, prosodic boundary tones are much more likely to succeed nouns than adjectives. This correlation can be exploited in many spoken language systems; in ASR, for example, knowledge of pitch accent events may help de-emphasize and eliminate word hypotheses that may otherwise compete closely with, or even overshadow, the truth.

A. Previous Work

To our knowledge, there has only been a handful of previous efforts at integrating symbolic prosody in ASR systems. A theoretical framework for integrating symbolic prosody within ASR is described in Ostendorf *et al.* [5]. Wang *et al.* [6] incorporated a four-class lexical stress model induced by acoustic–prosodic

Manuscript received February 04, 2008; revised July 05, 2008. Current version published December 12, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ruhi Sarikaya.

S. Ananthakrishnan was with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA. He is now with BBN Technologies, Cambridge, MA 02138 USA (e-mail: shankar.ananthakrishnan@gmail.com).

S. S. Narayanan is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: shri@sipi.usc.edu).

Digital Object Identifier 10.1109/TASL.2008.2005347

1558-7916/$25.00 © 2008 IEEE

features in a first pass Viterbi search to improve recognition performance over a baseline ASR (small vocabulary, conversational telephone speech, weather information domain). A WER reduction of 0.4% absolute (5.3% relative) over the baseline error rate of 7.6% was reported using this technique. The "ground truth" stress labels for training the lexical stress classifier were obtained from the segmental MFCC features using "stressed" and "unstressed" phoneme variants. Thus, the ground truth labels are not based on human annotation.

The work of Hasegawa-Johnson *et al.* [7] and Chen *et al.* [8] is perhaps most relevant to the present discussion. They incorporated symbolic prosody within ASR by training prosody-dependent phoneme acoustic models which were implemented as explicit-duration hidden Markov models (EDHMMs). An augmented feature set consisting of segmental MFCC features and acoustic–prosodic features based on F0 observations was used for training these models. This was used in conjunction with a prosody-enriched language model, implemented as a joint model of word tokens and prosody symbols. Using these models, they achieved a word error rate (WER) reduction of 1.5% absolute (6% relative) over a baseline error rate of 25.1% on the Boston University Radio News (BU-RNC) [9] task. In obtaining these results, they used 90% of the prosodically annotated data for training and the remaining 10% for testing.

In our previous work, we incorporated categorical prosody models within ASR in a decoupled fashion for N-best list rescoring [10], obtaining a 1.2% relative reduction in WER. On the prosody-integrated syllable recognition task [11], we rescored ASR lattices with categorical prosody models to obtain a 2% relative improvement in WER. These results were obtained on the BU-RNC.

B. Contributions of This Paper

While the above literature reports modest but significant improvement in word error rate with prosody-integrated ASR, one major limitation of categorical prosody models is the lack of sufficiently large prosody-transcribed data to train them. This is a direct result of the laborious and time-consuming nature of manual transcription of prosodic events. It has been reported [12] that obtaining human-annotated ToBI-style transcriptions for speech corpora can take up to 100–200 times real time, depending on the annotator's level of training and familiarity with the scheme. The difficulty and cost involved in obtaining such rich transcriptions is further evidenced by the existence of just a handful of publicly available corpora that provide ToBI transcriptions: the Boston University Radio News Corpus (BU-RNC) [9], the Boston Directions Corpus (BDC) [13], portions of the Rochester TRAINS Corpus [14], and a small subset of the Switchboard Corpus [15]. The unavailability of significant amounts of prosodically annotated data across different domains precludes, or at best severely limits the widespread adoption of categorical models of prosody for spoken language systems, including automatic speech recognition.

In this paper, we address the sparsity problem by proposing unsupervised adaptation of categorical prosody models using models trained from seed (human-annotated) data and a large speech corpus that does not provide enriched prosodic transcriptions. This not only improves the quality and coverage of the models, but also enables us to use them for larger scale tasks where the seed models trained from limited annotated data may be too impoverished to provide any benefit over a baseline system. Following our work in [16], we introduce novel techniques for adapting and smoothing the proposed categorical prosody models and evaluate their performance on two tasks: prosody label classification and speech recognition. To our knowledge, this is one of the first efforts at unsupervised adaptation of prosody models.

Our prosody-enriched ASR setup also differs from that of [7], [8] in that we do not modify the baseline ASR models in any way; as we will see, prosody models are used to rescore ASR-generated lattices in a postprocessing step. Thus, the prosody models are decoupled from the ASR models, allowing us to develop, adapt and evaluate them independently. This is quite different from the approach in [8], where prosody-dependent allophones are trained using an augmented feature set. This approach adds complexity to the ASR acoustic models and is not easy to scale to larger task domains for which prosody labels are not readily available for training.

A block diagram illustrating the various components of our system is shown in Fig. 1. The proposed method works with a relatively smaller amount of prosodically transcribed data for training. In order to evaluate our proposed adaptation techniques, we use a much smaller fraction ($<50\%$) of the human-annotated corpus for training (for comparison, [8] uses 90% of the prosody-annotated corpus for training and only 10% for evaluation).

The remainder of this paper is organized as follows. Section II describes the speech corpora we use in our work and presents details on the baseline speech recognizer. Section III introduces the architecture of the prosody-enriched ASR, and describes the role of the underlying prosodic acoustic and language models. We also discuss the implementation of this system: a lattice-enrichment procedure by which symbolic prosody is used to augment ASR lattices. In Section IV, we discuss the difficulty of training a prosodic language model (PLM) for prosodic event detection and speech recognition, and introduce a novel, unsupervised adaptation algorithm based on fractional n-gram counts derived from word-confusion networks. Section V describes a technique for improving the quality of the prosodic acoustic model (PAM) using a weighted variant of the expectation-maximization (EM) algorithm for maximum *a posteriori* (MAP) adaptation. Both adaptation techniques make use of a large speech corpus that, unlike the BU-RNC, is not annotated with symbolic prosody tags. In Section VI, we evaluate these adaptation schemes on the binary (presence versus absence) pitch accent classification task. We also evaluate the adapted prosody models in the context of speech recognition, and compare it with the baseline system that uses no prosody. Section VII concludes this paper with a discussion of our findings and proposes new directions for research in this area.

II. Speech Corpora and Baseline ASR

We used two publicly available speech corpora for our experiments in prosody model adaptation and enriched ASR; one with human-annotated ToBI transcriptions for training the seed prosody models, and the other, a standard corpus without

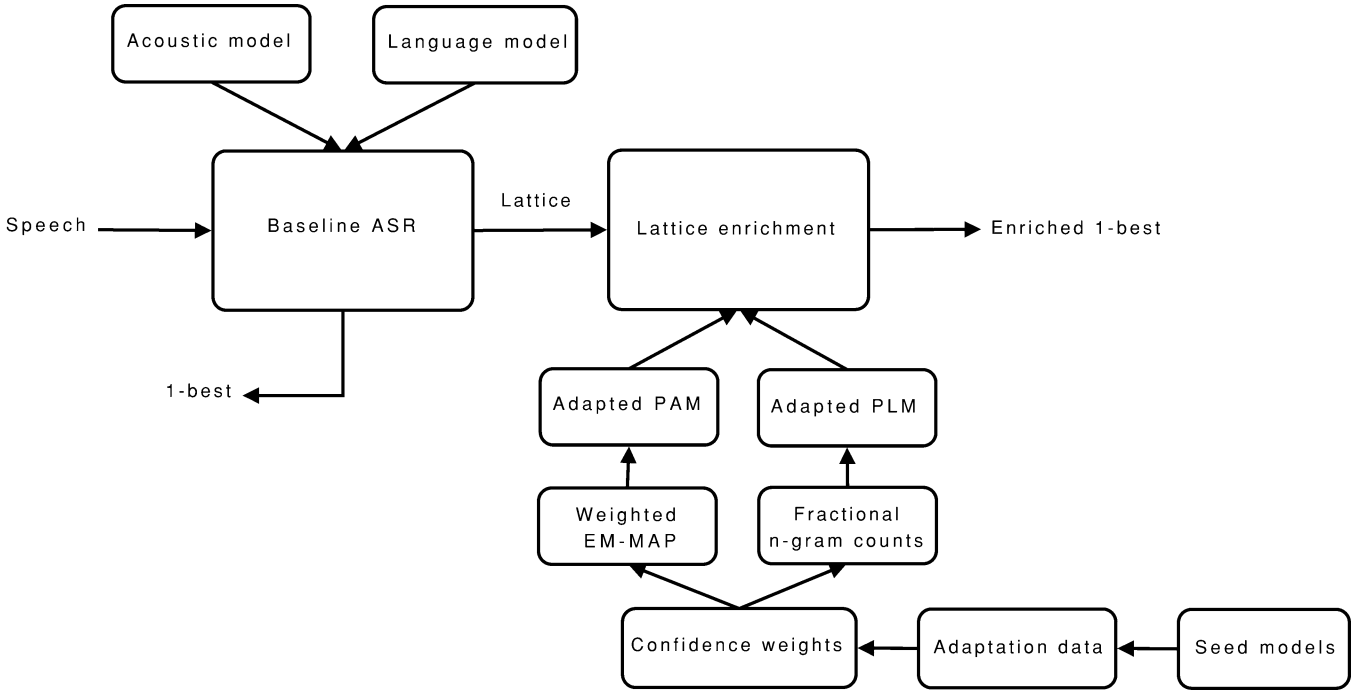

Fig. 1. Block diagram outlining the various components of the proposed system.

prosodic annotation for unsupervised adaptation of the seed models, as described below.

A. Human-Annotated Corpus

The Boston University Radio News Corpus (BU-RNC) [9] consists of about 3 h of read news broadcast speech from six speakers (three male, three female) with ToBI-style pitch accent and boundary tone annotations. The entire corpus consists of 29 573 words, which we split into a training set (14 719 words) and an evaluation set (14 854 words). After eliminating story repetitions from the evaluation set, its useful size was reduced to 10 273 words, which we split into a held-out development set (2900 words) and a test set (7373 words). A relatively large evaluation set was chosen in order to ensure that we were able to establish the statistical significance of our results. This also meant working with a smaller training set, which allowed us to evaluate our adaptation schemes in a data-starved scenario. As before [4], various types of pitch accents annotated in the BU-RNC were collapsed to binary labels that indicated presence or absence of pitch accents. A total of 7002 words (47.5%) in the training set carried any type of pitch accent; similarly, 3471 words (47.0%) in the test set carried a pitch accent. This corpus was used for training the seed prosodic acoustic and language models, and also for evaluating the performance of our unsupervised adaptation techniques. The prosody-enriched ASR was also evaluated on this corpus.

B. Adaptation Corpus

The adaptation dataset was culled from the WSJ1 (CSR-II) [17] newswire dictation corpus and consisted of approximately 22 400 utterances totaling about 407 000 words. This corpus consists of speech data and associated transcriptions, but does not provide symbolic transcriptions of pitch accents or other prosodic events. The unsupervised algorithms described in the following sections used this corpus to adapt the seed models for pitch accent classification and speech recognition.

C. Baseline Speech Recognizer

We developed the baseline ASR using the University of Colorado SONIC [18] continuous speech recognizer. We trained gender-specific acoustic models for this ASR by adapting seed acoustic models from the Wall Street Journal (WSJ) task with the training partition of the BU-RNC using tree-based MAPLR. The acoustic-phonetic features used by these models were standard 39-dimensional MFCC vectors used in most systems. The language model for the baseline ASR was trained with a combination of the BU-RNC training partition and the WSJ1 transcripts; the weight used to combine these two sources of text was optimized on the held-out development set. We used a trigram structure for the baseline LM and used the SRILM toolkit [19] to estimate its parameters.

III. Prosody-Enriched Speech Recognizer

The architecture of the prosody-enriched speech recognizer is developed along the lines of our previous work [10], [11], with a few important differences in the nature of the prosody models, as described below. We augment the standard ASR maximum *a posteriori* equation to include categorical prosody labels and acoustic–prosodic features as

$$\begin{aligned}(\mathbf{W}^*, \mathbf{P}^*) &= \arg\max_{\mathbf{W},\mathbf{P}} p(\mathbf{W}, \mathbf{P} \mid \mathbf{A_s}, \mathbf{A_p}) \\ &= \arg\max_{\mathbf{W},\mathbf{P}} p(\mathbf{W}, \mathbf{P}, \mathbf{A_s}, \mathbf{A_p}) \qquad (1)\end{aligned}$$

where $\mathbf{W}$ denotes the sequence of words, $\mathbf{P}$ the sequence of discrete prosody labels, $\mathbf{A_s}$ the segment-level spectral features

(MFCCs) and $\mathbf{A_p}$ the acoustic–prosodic features, typically consisting of energy and F0 contour statistics, as well as timing cues such as vowel duration. In order to make the problem more tractable, we simplify this joint probability distribution by invoking the following conditional independence assumptions.

- The prosody labels $\mathbf{P}$ are conditionally independent of the segment-level features $\mathbf{A_s}$ given the word sequence $\mathbf{W}$.
- The acoustic–prosodic features $\mathbf{A_p}$ are conditionally independent of the word sequence $\mathbf{W}$ and segment-level features $\mathbf{A_s}$ given the prosody labels $\mathbf{P}$.

Upon making these assumptions, (1) simplifies to

$$\begin{aligned}(\mathbf{W}^*, \mathbf{P}^*) &\\ \approx \arg\max_{\mathbf{W},\mathbf{P}} & \, p(\mathbf{W})p(\mathbf{A_s}\,|\,\mathbf{W})p(\mathbf{P}\,|\,\mathbf{W})p(\mathbf{A_p}\,|\,\mathbf{P})\\ \approx \arg\max_{\mathbf{W},\mathbf{P}} & \, \underbrace{p(\mathbf{A_s}\,|\,\mathbf{W})}_{\text{AM score}} \cdot \underbrace{p(\mathbf{A_p}\,|\,\mathbf{P})^{\alpha} p(\mathbf{W},\mathbf{P})^{\beta}}_{\text{prosody score}}. \qquad (2)\end{aligned}$$

Note that this formulation is different from that described in our previous work. In [10], we retained both the acoustic and language model scores provided by the ASR. In [11], we retained the ASR acoustic model score and replaced the ASR language model by a prosody-enriched factored language model $p(\mathbf{W}\,|\,\mathbf{P})$. In both [10] and [11], the prosodic acoustic model $p(\mathbf{P}\,|\,\mathbf{W})$ exhibited a discriminative structure, implemented as a feedforward neural network. In our present formulation, the prosodic acoustic model $p(\mathbf{A_p}\,|\,\mathbf{P})$ has a generative structure, while the prosodic language model $p(\mathbf{W},\mathbf{P})$ is now a joint distribution over compound tokens formed by the concatenation of the word sequence $\mathbf{W}$ and the corresponding pitch accent tags $\mathbf{P}$. The rationale for these choices is as follows.

- Experiments indicate that, while a factored back-off structure provides important smoothing benefits for predicting prosody labels [4], it is not as useful for predicting word sequences. In this case, a simpler structure based on trigrams of compound (word, prosody label) tokens is easier to train and provides greater flexibility.
- The choice of a generative structure for the prosodic acoustic model facilitates the conversion of lattices to confusion networks for generation of confidence weights for adaptation. Details will be explained in Section IV-A.

The weights α and β serve to weight the prosodic acoustic and language models for optimal performance and are tuned on the held-out development set. Below, we give details of the structure and implementation of the prosody models introduced in this section.

A. Prosodic Acoustic Model

The prosodic acoustic model in (2) has a generative structure, and a Gaussian mixture model (GMM) is a natural implementation for this model. Since the word-level pitch accent labels are binary (present versus absent), we trained two GMMs, one for each class, using the standard EM algorithm for GMM parameter estimation. Word-level acoustic–prosodic features for training these GMMs are obtained from ASR forced alignment at the word- and phone-level, and are based on previous work on prosody labeling [4]; the feature set was expanded to incorporate surrounding context information (previous and succeeding words) in addition to features drawn from the current word. We extracted a total of ten features related to the F0 track, RMS energy, and vowel duration cues as described in Table I. We used the normalized duration of the longest vowel within the word as a durational feature. F0-related features include the ratio of within-word average and maximum F0 to the corresponding statistics extracted from the preceding and succeeding words, the ratio of within-word average F0 to the utterance average F0, and within-word F0 range. Energy-related features include the ratio of the within-word RMS energy to the utterance average RMS energy, as well as the ratio of within-word RMS energy to the RMS energy of the preceding and succeeding words.

TABLE I
ACOUSTIC–PROSODIC FEATURES

Feature	Description
VOWEL_DUR	$\max_{v \in w_i} norm\_dur(v)$
F0AVG_UTT	$\lvert avgF0(w_i) - avgF0(utt)\rvert$
F0RANGE	$maxF0(w_i) - minF0(w_i)$
F0AVG_PAVG	$\lvert avgF0(w_i) - avgF0(w_{i-1})\rvert$
F0AVG_NAVG	$\lvert avgF0(w_i) - avgF0(w_{i+1})\rvert$
F0MAX_PMAX	$\lvert maxF0(w_i) - maxF0(w_{i-1})\rvert$
F0MAX_NMAX	$\lvert maxF0(w_i) - maxF0(w_{i+1})\rvert$
ERMS_AVG	$rmse(w_i)/rmse(utt)$
ERMS_PRMS	$rmse(w_i)/rmse(w_{i-1})$
ERMS_NRMS	$rmse(w_i)/rmse(w_{i+1})$

B. Prosodic Language Model

The PLM is a joint probability distribution over the word sequence $\mathbf{W}$ and binary pitch accent tags $\mathbf{P}$. We implemented it by creating compound tokens $\mathbf{W}' = (\mathbf{W},\mathbf{P})$ and training a standard back-off trigram LM with these tokens. This structure was chosen because of its simplicity, efficiency of implementation, and close relationship with the standard ASR LM. This model is trained only on the annotated data from the BU-RNC. We used the SRILM toolkit to train this model.

C. Lattice Enrichment

The baseline ASR can be configured to generate lattices that provide a compact representation of many competing word hypotheses. Each lattice typically consists of a number of nodes, which mark points in time, and arcs which connect these nodes. Each arc represents one word and carries supplementary information including the ASR acoustic model score for that word, the language model score for that word given its context, and a phone-level segmentation of the word.

For each word in the lattice, we extract the acoustic–prosodic features listed in Table I using the word- and phone-level time-alignment information. One issue is that these lattices can be quite dense, and it may not be possible to establish a unique left- and right-context for feature extraction (since feature values for the current word depend on statistics drawn not only from the current word, but also from the previous and succeeding words). Therefore, we first expand the lattices so as to establish a unique left- and right-context for each arc (word) in the lattice. We then extract the features listed above and evaluate their likelihood given the prosodic acoustic model $p(\mathbf{A_p}\,|\,\mathbf{P})$.

The two likelihood scores per lattice arc (one from each GMM) are embedded in the ASR lattices as illustrated in Fig. 2 to generate prosody-enriched lattices. The procedure

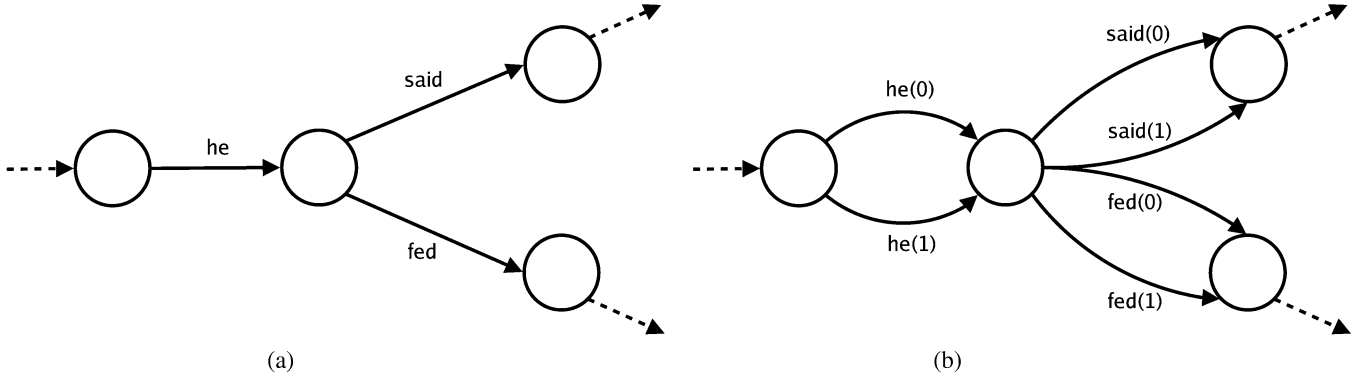

Fig. 2. Plain ASR-generated baseline lattice on the left. Prosody-enriched equivalent on the right. (a) Plain lattice. (b) Enriched lattice.

can be described as follows. We replace each word arc in the original ASR lattice with its two prosodic variants, one with a pitch accent and the other without. The ASR acoustic model score associated with the original arc is then applied to both replacement arcs after combining it with the appropriate likelihood score from the prosodic acoustic model. Scores from the prosodic acoustic model are weighted as per (2) before they are combined with the ASR acoustic model score. ASR language model scores are left untouched during this step.

In the next step, the ASR language model scores $p(\mathbf{W})$ associated with each word arc in the prosody-enriched lattices are replaced with scores $p(\mathbf{W},\mathbf{P})$ from the prosodic language model with the appropriate weighting factor applied. Following the description in (2), the resulting enriched lattices contain a combination of the ASR acoustic model score $p(\mathbf{A_s}\,|\,\mathbf{W})$, the prosodic acoustic model score $p(\mathbf{A_p}\,|\,\mathbf{P})$, and the prosodic language model score $p(\mathbf{W},\mathbf{P})$. Finally, (2) is implemented by using the Viterbi algorithm to find the most likely path through these enriched lattices.

IV. Prosodic Language Model Adaptation

We note from the implementation of the prosody-enriched ASR in Section III-C that the ASR LM is replaced by the prosodic LM. One major issue with this substitution is data sparsity. The ASR LM $p(\mathbf{W})$ is trained on very large text corpora with millions of words (although in our experiments, it has been trained with ca. 415 000 words for a fair comparison between the baseline and enriched systems). On the other hand, prosodically annotated text required to train the PLM $p(\mathbf{W},\mathbf{P})$ is in very short supply: the training set used in our experiments contains only 14 719 words with pitch accent annotations for training this model. If a PLM trained with such a small dataset replaces the baseline ASR LM as proposed, the performance of the enriched ASR will be very poor due to the impoverished PLM's lack of coverage.

In order to address sparsity and coverage issues endemic to the seed PLM, we present a novel technique for unsupervised adaptation of the seed PLM using a large, prosodically unlabeled speech corpus. We evaluate the effectiveness of our method on the binary pitch accent (presence versus absence) classification task before integrating the adapted PLM within a speech recognizer. Our technique works by weighting different segments of the adaptation set according to the confidence level assigned by the seed prosody models during automatic prosody labeling. This weighted data is used to adapt the seed PLM. Details of this method are presented below. For the task of prosodic event detection, we will assume throughout this section that we have access to the speech data as well as the corresponding clean text transcriptions; this data is provided by the adaptation set.

A. Architecture of the Prosodic Event Classifier

The prosodic event classifier used in these experiments follows our work in [4], where we proposed a MAP structure for the prosody recognizer. Thus, our system chooses the sequence of binary pitch accent labels $\mathbf{P}$ that maximizes their posterior probability given the acoustic–prosodic features $\mathbf{A_p}$ and the word sequence $\mathbf{W}$ as

$$\mathbf{P}^* = \arg\max_{\mathbf{P}} p(\mathbf{P}\,|\,\mathbf{A_p},\mathbf{W}). \tag{3}$$

The above expression can be simplified by first applying Bayes' rule and then by invoking the assumption that the acoustic–prosodic features are conditionally independent of the lexical evidence, given the sequence of pitch accent labels. We can then rewrite (3) as follows:

$$\begin{aligned}\mathbf{P}^* &= \arg\max_{\mathbf{P}} p(\mathbf{A_p},\mathbf{W}\,|\,\mathbf{P})p(\mathbf{P})\\ &\approx \arg\max_{\mathbf{P}} p(\mathbf{A_p}\,|\,\mathbf{P})^{\gamma} p(\mathbf{W},\mathbf{P}).\end{aligned} \tag{4}$$

In (4), the RHS involves two factors—1) the prosodic acoustic model $p(\mathbf{A_p}\,|\,\mathbf{P})$, which provides the likelihood of the acoustic–prosodic features given the pitch accent label and 2) the PLM $p(\mathbf{W},\mathbf{P})$, which relates the word sequence to the pitch accent label sequence. We note that the models introduced in this section for automatic prosodic event detection are congruous to those used in our proposed enriched speech recognizer.

B. Implementation of the Classifier

Implementation of the classifier in (4) begins with the construction of a word confusion network that encodes the prosodic variants of each word in the utterance to be labeled. This is illustrated in Fig. 3, where a linear sequence of nodes is connected by pairs of arcs; in the figure, the top arc denotes the unaccented variant of the word, while the lower arc denotes the

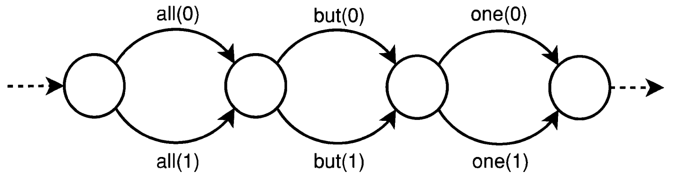

Fig. 3. Word graph with prosodic variants.

accented variant. ASR-generated word- and phone-level alignments are used to derive the acoustic–prosodic features listed in Table I. The prosodic acoustic model is then used to evaluate likelihood scores for each prosodic variant; these scores are embedded within the appropriate arc of the confusion network. This network is then rescored with the PLM, and the resulting lattice is decoded with the Viterbi algorithm to determine the most likely sequence of pitch accent labels. The mixing weight γ is optimized on the held-out development set.

Note that it is also possible to use the acoustic and language models individually for prosodic event detection. To use the former by itself, we use the GMMs directly as a MAP classifier by multiplying the model likelihood scores with the prior distribution of labels $p(\mathbf{P})$, disregarding the PLM rescoring step. To use the PLM by itself, we simply score the original confusion network (omitting the acoustic likelihood scores) with the PLM.

We used the small, human-annotated BU-RNC training partition (14 719 words) to train the seed prosodic acoustic and language models. For bootstrapping, we implemented the prosodic acoustic model as 25-mixture GMMs with diagonal covariance structure. The seed PLM was implemented as a standard back-off trigram model and was trained with compound tokens consisting of the word sequence and the corresponding pitch accent label sequence derived from ToBI annotation. As expected, the PLM exhibited very poor coverage, with a 21.3% token out-of-vocabulary (OOV) rate on the test set. Improving the quality and coverage of the PLM is therefore of significant importance for both prosodic event labeling and for prosody-enriched ASR; we achieve this through unsupervised adaptation.

C. Naïve Adaptation

Before presenting our proposed scheme, we describe two naïve techniques for alleviating the high OOV rate of the seed model and evaluate their performance with respect to the seed model as well as the proposed adaptation technique.

- *Majority Prediction (MajPred)*: This is a trivially straightforward method in which we associate each word of the adaptation set with the majority prosodic category. In our case, each word is labeled as "not accented," and this labeled corpus is then used to train a PLM that is merged with the seed model to create an adapted model $p_{\mathrm{all0}}(\mathbf{W}')$.
- *LM Prediction (LMPred)*: In this method, we train a factored back-off model $p_{\mathrm{seed}}(\mathbf{P} \mid \mathbf{W})$ from the annotated data. The factored structure provides better smoothing for prosody label prediction as compared to the joint model and is described in greater detail in [4]. We use this model to predict prosody labels for the adaptation set. This labeled corpus is then used to train a PLM that is merged with the seed model to generate the adapted model $p_{\mathrm{lmp}}(\mathbf{W}')$.

In both cases, the weight used to merge the seed model and the model trained from the adaptation data is optimized on the held-out development set. Neither of these simplistic techniques utilizes the discriminatory power of the acoustic evidence. As we will see, our proposed scheme makes use of both acoustic and lexical models of prosody to train the PLM.

D. Proposed Adaptation Scheme

We begin by setting up the pitch accent detection framework for the unlabeled adaptation data using the acoustic model and the seed PLM as described in Section IV-A. Due to the back-off structure of the PLM, the lattices generated by rescoring the word graph with the seed models no longer retain the original sausage structure.

In the next step, we generate word posterior probabilities for each compound token $W' = (W, P)$. This is accomplished by a two-step process: 1) link posteriors $p(l \mid \mathbf{A}_{\mathbf{P}})$ are computed for each link l in the rescored lattice using a variant of the forward-backward algorithm, and 2) links corresponding to the same compound token are collapsed to generate a confusion network identical to the one that was originally created for labeling, except that the arcs in the network now contain compound token posterior probabilities computed from the prosodic acoustic and language models. This technique for generating word posteriors and confusion networks is borrowed from minimum word error rate decoding for ASR [20], [21].

The confusion networks with token posteriors are then used to generate fractional (soft) counts for PLM n-gram estimation. Fig. 4 illustrates this procedure for a sample fragment of the network. The unigram count of a compound token is set to its posterior probability. For higher order terms (bigrams and trigrams), the count of the term is set to the geometric mean of its constituent posteriors. Thus, terms with high posterior probabilities, which correspond to regions in which the seed models exhibit high confidence, are assigned larger fractional counts and vice-versa. In this way, fractional counts for all possible unigrams, bigrams, and trigrams are extracted from the adaptation networks and are used to train a PLM $p_{\mathrm{frac}}(\mathbf{W}')$ from the adaptation data. This is merged with the seed PLM to create the adapted model $p_{\mathrm{adapt}}(\mathbf{W}')$. As before, the merge weight is optimized on the held-out development set.

We defer the presentation and discussion of results for binary pitch accent (presence versus absence) classification with the adapted prosodic language models (both naïve and proposed techniques) to Section VI.

V. Prosodic Acoustic Model Adaptation

While the parameter-rich prosodic language model bears the brunt of the data sparsity issue discussed earlier in this paper, the problem also affects the prosodic acoustic model, but to a lesser degree, depending on the dimensionality of the acoustic–prosodic features and the complexity of the constituent GMMs. In our case, a GMM may require several hundred free parameters to be trained with just a few thousand training samples, causing the model to overfit the training set and prevent generalization to unseen data.

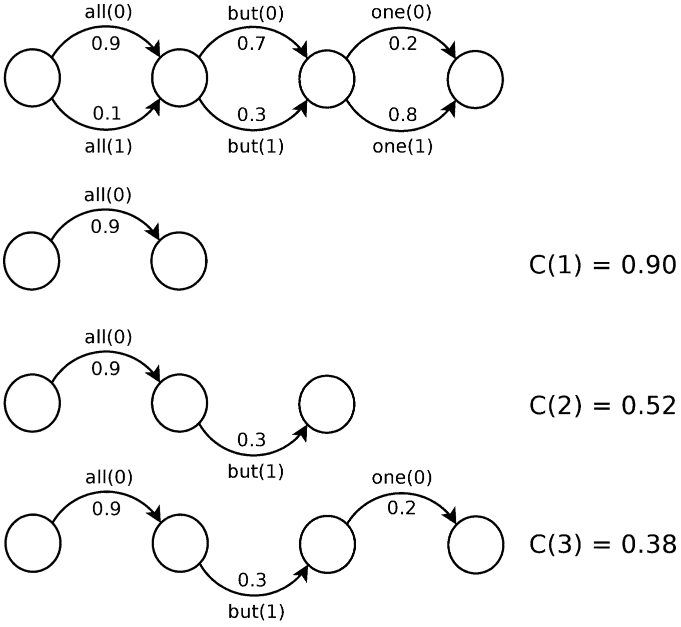

Fig. 4. Obtaining n-gram counts from lattice chunks.

In this section, we present a technique for unsupervised adaptation of GMM-based prosodic acoustic models using a much larger, unannotated dataset. The first part of this process, which involves obtaining confidence weights (posterior probabilites) for each compond token, is identical to the procedure followed for PLM adaptation. This is followed by MAP adaptation of the seed prosodic acoustic models using a weighted variant of the EM algorithm. Just as we did for the PLM adaptation case, the testbed for evaluating our proposed technique is the binary pitch accent classification problem. We show that the adapted models outperform the seed models on this task.

A. Acoustic Model Structure

In Section IV-B, we chose a relatively low-complexity 25-mixture diagonal covariance GMM structure for the acoustic model and used it to generate confidence weights for PLM adaptation. While such a model could be robustly estimated with the small number of training samples at our disposal, a more complex model with full-covariance structure is likely to improve classification and recognition performance given a sufficiently large training set for estimating its parameters. We therefore chose 45-mixture, full-covariance GMMs for ultimately implementing the pitch accent classifier and prosody-enriched ASR. However, we were unable to train these models with the limited training data due to numerical errors caused by ill-conditioned covariance matrices, a problem that is typical of such scenarios. Our approach was to train 45-mixture diagonal-covariance models and use their parameters as initial "guess" values for unsupervised adaptation, which uses the unlabeled dataset to estimate full-covariance models.

B. Weighted EM-MAP

We propose a novel, weighted EM-MAP scheme for soft adaptation of the acoustic–prosodic models using posterior probabilities obtained from the prosodic confusion networks. This method differs from conventional EM training for GMM estimation in that each adaptation sample has a weight associated with it. Samples with larger weights (indicative of high confidence) contribute more to the adaptation process, whereas samples with low confidence do not have a significant influence on the adapted estimates. A distinct feature of this approach is that we do not divide the unlabeled data into classes based on confidence scores; rather, all adaptation samples affect both acoustic–prosodic models simultaneously, but to different degrees. The relative influence of each sample on the GMMs is dictated by the external information source, in this case the posterior probability assigned to each sample by the seed models.

We begin by defining a likelihood function that incorporates the seed model weights as

$$L(\boldsymbol{\Theta} \,|\, \mathbf{X}, \mathbf{B}) = p(\mathbf{X} \,|\, \boldsymbol{\Theta}, \mathbf{B}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \omega_k p_k(x_i \,|\, \theta_k, \beta_i) \tag{5}$$

where $p_k(x_i \,|\, \theta_k, \beta_i) \equiv \mathcal{N}(x_i; \mu_k, \beta_i^{-1}\Sigma_k)$. The full set of model parameters is represented by the vector $\boldsymbol{\Theta}$. This function differs from the traditional likelihood function due to integration of the confidence weights $\mathbf{B} = \{\beta_1, \ldots, \beta_N\}$ associated with vector adaptation samples $\mathbf{X} = \{x_1, \ldots, x_N\}$. The rationale behind this modified likelihood function is that adaptation samples associated with a large weight see a narrow, focused distribution, whereas samples with low confidence weights see a diffuse, flat distribution. This formulation leads to parameter update equations that emphasize samples with high confidence and vice-versa.

Following the notation of [22], the modified auxiliary function for EM is then given as (the superscript in $\boldsymbol{\Theta}^{\mathbf{g}}$ indicates an initial "guess" for the parameter $\boldsymbol{\Theta}$).

$$\begin{aligned} Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{\mathbf{g}}) &= \mathbb{E}(\log p(\mathbf{X}, \mathbf{Y} \,|\, \boldsymbol{\Theta}, \mathbf{B}) \,|\, \mathbf{X}, \boldsymbol{\Theta}^{\mathbf{g}}, \mathbf{B}) \\ &= \sum_{\mathbf{y} \in \mathcal{Y}} \log p(\mathbf{X}, \mathbf{Y} \,|\, \boldsymbol{\Theta}, \mathbf{B}) p(\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\Theta}^{\mathbf{g}}, \mathbf{B}) \\ &= \sum_{k=1}^{K} \sum_{i=1}^{N} c_{ik} \log(\omega_k p_k(x_i \,|\, \theta_k, \beta_i) \end{aligned} \tag{6}$$

where

$$c_{ik} = p(k \,|x_i, \boldsymbol{\Theta}^{\mathbf{g}}, \beta_i) = \frac{\omega_k^g p_k(x_i \,|\, \theta_k^g, \beta_i)}{\sum_{l=1}^{K} \omega_l^g p_l\,(x_i \,|\, \theta_l^g, \beta_i)}.$$

In the above formulation, $\boldsymbol{\Theta}$ represents the full set of GMM parameters, including mixture weights, mean vectors, and covariance matrices. The vector $\mathbf{Y}$ represents the (hidden) random vector that determines the assignment of training samples to mixture components. Its dimensionality is equal to the number of training samples in $\mathbf{X}$.

Using basic vector and matrix calculus [22], this modified auxiliary function can be maximized w.r.t the unknown parameters to obtain the following maximum-likelihood (ML) update

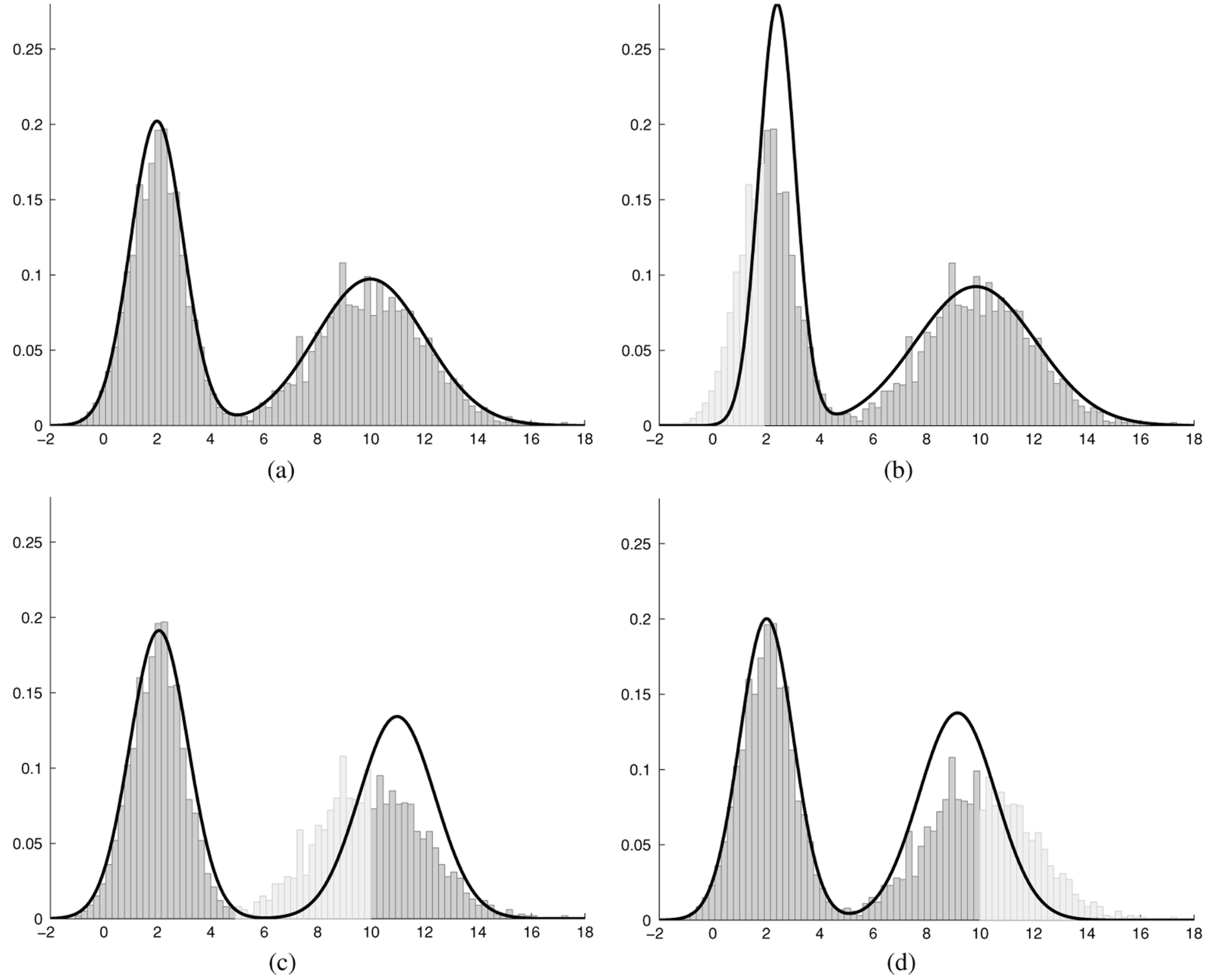

Fig. 5. One-dimensional example illustrating weighted EM-MLE. Dark regions of the histogram represent data samples with unit weight. Lighter shaded regions of the histogram represent data samples with a uniform lower weight of 0.3. (a) Unweighted dataset. (b) Weighted set I. (c) Weighted set II. (d) Weighted set III.

equations for the mixture weights ω_k, mean vectors μ_k, and covariance matrices Σ_k as follows:

$$\omega_k^{'} = \frac{1}{N}\sum_{i=1}^{N} c_{ik} \tag{7}$$

$$\mu_k^{'} = \frac{\sum_{i=1}^{N}\beta_i c_{ik} x_i}{\sum_{i=1}^{N}\beta_i c_{ik}} \tag{8}$$

$$\Sigma_k^{'} = \frac{\sum_{i=1}^{N}\beta_i c_{ik}(x_i-\mu_k^{'})(x_i-\mu_k^{'})^t}{\sum_{i=1}^{N} c_{ik}}. \tag{9}$$

These modified update equations make intuitive sense: (8) is the mean of the adaptation samples weighted not only by the mixture occupation probability c_{ik} as in conventional EM, but also by the confidence weights β_i. This suggests that adaptation samples with large confidence weights influence the estimated GMM mean vectors to a greater extent than samples with low weights. Similarly, (9) implies that the distributions are focused around samples with large confidence weights.

Fig. 5 illustrates the operation of these equations for a one-dimensional dataset over which we fit a two-mixture GMM. Fig. 5(a) shows, superimposed on a histogram of the dataset, a trace of the GMM estimated using the standard EM algorithm without weights. Fig. 5(b)–(d) illustrates the GMMs estimated by weighted EM-MLE for different cases; in each of these cases, unit weight is applied to the samples with the darker shade, while a lower weight (0.3) is applied to the samples with the lighter shade. These figures clearly show that the weighted EM-MLE equations exhibit the desired behavior of biasing the GMM towards the samples with larger weight.

While the ML update equations provide intuition on how the confidence weights impact parameter estimation, our task in this paper is to adapt existing seed acoustic–prosodic models using unlabeled data. MAP adaptation is the traditional approach to this problem. Following the approach of [23], we construct a prior distribution for the GMM parameters by assuming the form of a Dirichlet distribution for the mixture weights ω_k and a normal-Wishart distribution for the mean vectors μ_k and covariance matrices Σ_k

$$P(\mathbf{\Theta}) \propto \prod_{k=1}^{K} \omega_k^{\lambda_k} |\Sigma_k^{-1}|^{\alpha_k - d/2} \cdot \exp\left(-\frac{\tau_k}{2}(\mu_k - m_k)^t \Sigma_k^{-1} (\mu_k - m_k)\right) \cdot \exp\left(-\mathrm{tr}\left(U_k \Sigma_k^{-1}\right)\right). \tag{10}$$

The prior "hyperparameters" $\lambda_k, \alpha_k, \tau_k, m_k$ and U_k are computed using the original (labeled) seed training data in a manner similar to that described in [24]. This leads to the following update equations for weighted EM-MAP:

$$\omega_k^{'} = \frac{\lambda_k + \sum_{i=1}^{N} c_{ik}}{N + \sum_{k=1}^{K} \lambda_k} \tag{11}$$

$$\mu_k^{'} = \frac{\tau_k m_k + \sum_{i=1}^{N} \beta_i c_{ik} x_i}{\tau_k + \sum_{i=1}^{N} \beta_i c_{ik}} \tag{12}$$

$$\Sigma_k^{'} = \frac{2U_k + S_k + M_k}{2\alpha_k - d + \sum_{i=1}^{N} c_{ik}} \tag{13}$$

where, for ease of notation, we have defined S_k and M_k as follows:

$$S_k = \sum_{i=1}^{N} \beta_i c_{ik} (x_i - \mu_k^{'})(x_i - \mu_k^{'})^t \tag{14}$$

$$M_k = \tau_k (m_k - \mu_k^{'})(m_k - \mu_k^{'})^t. \tag{15}$$

As with standard EM, (11), (12), and (13) are evaluated iteratively until convergence.

VI. Experimental Results

We began by training seed prosodic acoustic and language models from the training partition of the BU-RNC. After evaluating the performance of these models on the binary pitch accent classification task in order to establish a baseline, we adapted these models using the proposed techniques. The adapted models were then used to carry out the same classification task. Results from these experiments are summarized below.

A. Prosodic Language Model Adaptation

The seed PLM trained from human-annotated BU-RNC data gave an error rate of 32.0% on the test data, significantly worse than the seed acoustic model. This is on account of the high compound token OOV rate of 21.3%, which is a result of data sparsity. We then experimented with the two naïve methods discussed in Section IV-C to generate adapted PLMs and evaluated their performance. The majority prediction (MajPred) scheme performed worse than the seed PLM with an error rate of 34.5%. The token OOV rate was significantly reduced to 14.8% due to inclusion of the adaptation data, but since the two prosodic categories were more or less balanced, majority prediction resulted in a high error rate when used to label the adaptation set, causing poor estimates of the resulting PLM. The LM prediction scheme (LMPred), with an error rate of 31.6% and a token OOV rate of 12.8%, performed slightly better than the seed model. The incremental gain provided by this scheme is due to improved smoothing provided by the factored PLM.

We then implemented our proposed confidence-weight based technique of Section IV-D to generate the adapted model $p_{\text{adapt}}(\mathbf{W}')$. The weight of the acoustic model in the initial rescoring step and the PLM merge weight in the final step were jointly optimized on the development set. The error rate for pitch accent labeling with this model was 27.6%, which represented a 4.4% absolute (13.8% relative) reduction in error rate over the seed model. The token OOV rate for this model was only 4.8%, a 16.5% absolute reduction over the seed model. These results are summarized in Table II.

TABLE II
Pitch Accent Detection Results (Error Rate)

Method	Dev	Test	OOV
Chance	50.0%	47.0%	-
Acoustic	26.4%	26.6%	-
Seed PLM	31.0%	32.0%	21.3%
MajPred PLM	34.0%	34.5%	14.8%
LMPred PLM	29.7%	31.6%	12.8%
Adapted PLM	**24.1%**	**27.6%**	**4.8%**

B. Prosodic Acoustic Model Adaptation

To evaluate the performance of the adapted prosodic acoustic model, we divided the evaluation portion of the BU-RNC dataset into ten held-out development and cross-validation test sets with 90% of evaluation data (9246 samples) in the former and 10% (1027 samples) in the latter. The ten cross-validation test sets were independent of one another. In the first step, seed acoustic models with 25-mixtures and diagonal covariance were combined with the seed PLM to obtain confidence weights for the adaptation samples. With these weights and 45-mixture diagonal covariance models as a starting point, we implemented the weighted EM-MAP technique to obtain 45-mixture full-covariance models, which we used for classification. When implementing weighted EM-MAP, we ensured that samples corresponding to tokens not present in the PLM (OOV terms) were discarded so that the confidence weights contained contributions from both the prosodic acoustic and language models. The raw confidence weights were pruned so that only those samples with a large difference between the two class posteriors would be used for adaptation (these samples have a very high likelihood of being labeled correctly by the seed models). The pruned confidence scores were used to adapt the seed models. Two parameters were sequentially optimized by evaluating classification performance of the adapted models on the held-out development data: 1) the weight of the acoustic–prosodic model in (2) and 2) the pruning threshold for selecting samples from the adaptation set.

Table III summarizes pitch accent classification error performance of the seed and adapted models averaged across the ten cross-validation sets. It is clear that increasing the number of mixtures from 25 to 45, while maintaining the diagonal covariance structure does not improve classification performance significantly (only 0.5% relative on the test sets). On the other hand, we note that the adapted models reduce the classification

TABLE III
PITCH ACCENT CLASSIFICATION ERROR

Model	Dev	Test
Seed 1 (25 / diag)	27.0%	26.5%
Seed 2 (45 / diag)	26.4%	26.4%
Adapted (45 / full)	**25.2%**	**25.3%**

error rate by 4.5% relative to the 45-mixture seed models on the held-out sets ($p \leq 0.002$) and by 4.3% relative on the test sets ($p \leq 0.06$). We used the Wilcoxon matched-pairs signed-rank test to evaluate the statistical significance of these results.

C. *Prosody-Enriched Speech Recognizer*

We began by implementing the baseline ASR using the SONIC continuous speech recognizer as described in Section II-C. The language model weight was optimized on the held-out development partition; the WER of this system was 29.8% on the held-out set and 32.4% on the test partition. We also evaluated the oracle error rate on these sets by aligning the reference transcriptions with all possible lattice word hypotheses; these were determined to be 24.1% for the development set and 27.2% for the test set. These figures indicate that there is a 5.2% margin for improvement on the test set. The baseline WER for this domain is relatively high due to the following factors.

- We used a relatively small fraction ($<$50%) of the corpus with an in-domain data size of just 14 719 words for training the ASR. The rationale for this choice was to keep the ASR data sets consistent with the partitions used for prosodic event classification. This allowed us to use the same adapted prosody models for pitch accent classification as well as prosody-enriched ASR. Another reason for choosing a small training corpus was to demonstrate the efficacy of the proposed adaptation techniques in a sparse data scenario.
- We eliminated all story repetitions between the training and testing partitions. A few news stories in the corpus are read by more than one speaker; if we allowed the same story to exist in the training and testing sets, the baseline WER was significantly reduced because the language model was able to predict these stories extremely well. In fact, the overall WER on the 14 854 word evaluation set was just 23.3%, much lower than the figures quoted in this section. However, we chose to make the task more challenging by not permitting the same story to exist in the training and evaluation sets; this also lets us assess the value of prosody models for ASR.

Finally, we implemented the proposed method by using the adapted prosody models to enrich the baseline ASR lattices with categorical pitch accent tags as detailed in Section III-C. We performed Viterbi decoding on the enriched lattices to implement the augmented ASR of (2). The mixing parameters α and β were optimized on the development set. The error rate of the prosody-enriched ASR was determined to be 29.4% on the development set and 31.4% on the test set. Thus, the augmented ASR gives an absolute WER reduction of 1% (3.1% relative) over the baseline system. We used the NIST Matched Pairs Sentence Segment Word Error (MAPSSWE) test to establish that this reduction in error rate is statistically significant at the $p \leq 0.006$ level. These results are summarized in Table IV.

TABLE IV
ASR WORD ERROR RATES

System	Dev	Test
Baseline	29.8%	32.4%
Lattice oracle	24.1%	27.2%
Enriched	**29.4%**	**31.4%**

VII. DISCUSSION AND FUTURE WORK

We began this paper by introducing an architecture for augmenting automatic speech recognition systems with categorical prosody models. The acoustic and language models that make up the prosodic subsystem are decoupled from the baseline ASR and can be implemented and trained independently. We also presented a method for combining these models with the baseline ASR system through lattice enrichment.

One major limitation of categorical prosody models is the lack of sufficiently large speech corpora annotated with symbolic prosody labels for training said models. Data sparsity can result in poor estimates of model parameters. In case of the prosodic language model, it can also result in an unacceptably high OOV rate on the test set. In order to alleviate these issues, we introduced novel adaptation techniques for both prosodic acoustic and language models in an unsupervised setting using a large, unlabeled speech corpus. In both cases, we used the binary pitch accent classification testbed to generate confidence weights for unsupervised adaptation; the same task was also used to evaluate the adapted models.

To adapt the prosodic language model, we used confidence weights generated by the seed models to obtain fractional counts for compound tokens from the adaptation dataset. These counts were used to estimate a new language model which was merged with the seed model to obtain adapted model estimates. The proposed scheme resulted in a 13.8% relative reduction in binary pitch accent classification error rate and a significant reduction in token OOV rate over the seed model. This was found to be far superior to majority prediction and LM prediction, two naïve adaptation schemes to which we compared our proposed method. We note that, with the existing tools, it is only possible to train standard backoff n-gram models from the fractional counts, and not factored models; additionally, the tools do not yet permit two factored models to be merged. However, previous work on prosodic event classification [4] demonstrates the power of factored models for that task. On the other hand, a factored model is less useful as far as the prosody-enriched ASR is concerned. Therefore, this limitation may be viewed as a drawback for binary pitch accent classification, but not for prosody-augmented speech recognition. The increased flexibility of being able to generate a factored model as the output of our adaptation scheme is motivation enough for us to include it on our agenda for future work.

We also proposed a weighted EM-MAP algorithm for adaptation of the prosodic acoustic model. This is a more general version of standard MAP, in which each adaptation sample has an associated weight that indicates the "level of belongingness" of that sample to the model it is being used to adapt. These weights are derived from the adaptation dataset using the same pitch accent classification testbed with which we generated fractional counts for PLM adaptation. The proposed method allowed us to train full-covariance models (which we were unable to do with the seed data), and provided classification error reduction of 4.3% relative to the seed models on the binary pitch accent classification task. We wish to emphasize that the weighted EM-MLE and EM-MAP formulations introduced in this paper are quite general and may be applied to arbitrary data ranked by a knowledge source in order of their "importance" or "belongingness" through the assignment of numerical weights or scores. We also note that our proposed scheme does not employ any discriminative adaptation techniques. Thus, while adaptation will likely improve model fit and generalization, it does not necessarily guarantee better classification performance. This is a limitation for tasks such as pitch accent classification; however, it is less of a concern for prosody-enriched ASR. Indeed, MAP adaptation of ASR acoustic models (HMMs) is performed in a generative fashion [25].

Finally, we integrated the adapted models in the proposed prosody-enriched ASR framework. The 1-best test set WER of the baseline ASR without enrichment was 32.4%, whereas the lattice oracle error rate was determined to be 27.2%. An actual relative gain of 3.1% was obtained upon incorporation of the prosody models. These results demonstrate the usefulness of categorical prosody models in the context of speech recognition. An added advantage of the proposed enriched ASR is that we are able to obtain the sequence of pitch accent labels that correspond to the hypothesized word sequence.

References

[1] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard scheme for labeling prosody," in *Proc. Int. Conf. Spoken Lang. Process.*, 1992, pp. 867–869.

[2] D. Hirst and A. D. Cristo, , D. Hirst and A. D. Cristo, Eds., *Intonation Systems: A Survey of Twenty Languages*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[3] E. Grabe, F. Nolan, and K. Farrar, "IViE—A comparative transcription system for intonational variation in English," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, pp. 1259–1262.

[4] S. Ananthakrishnan and S. Narayanan, "Automatic prosodic event detection using acoustic, lexical, and syntactic evidence," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 216–228, Jan. 2008.

[5] M. Ostendorf, I. Shafran, and R. Bates, "Prosody models for conversational speech recognition," in *Proc. 2nd Plenary Meeting Symp. Prosody and Speech Process.*, 2003, pp. 147–154.

[6] C. Wang and S. Seneff, "Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 2761–2764.

[7] M. Hasegawa-Johnson, J. Cole, C. Shih, K. Chen, A. Cohen, S. Chavarria, H. Kim, T. Yoon, S. Borys, and J.-Y. Choi, "Speech recognition models of the interdependence among syntax, prosody and segmental acoustics," in *Proc. HLT/NAACL*, 2004, pp. 56–63.

[8] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S.-S. Kim, J. Cole, and J.-Y. Choi, "Prosody dependent speech recognition on radio news corpus of American english," *IEEE Trans. Speech, Audio, Lang. Process.*, vol. 14, no. 1, pp. 232–245, Jan. 2006.

[9] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Boston Univ., Boston, MA, Tech. Rep. ECS-95-001, Mar. 1995.

[10] S. Ananthakrishnan and S. Narayanan, "Improved speech recognition using acoustic and lexical correlates of pitch accent in a N-best rescoring framework," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2007, pp. 873–876.

[11] S. Ananthakrishnan and S. Narayanan, "Prosody-enriched lattices for improved syllable recognition," in *Proc. Int. Conf. Spoken Lang. Process.*, Antwerp, Belgium, Sep. 2007.

[12] A. Syrdal, J. Hirschberg, J. McGory, and M. Beckman, "Automatic ToBI prediction and alignment to speed manual labeling of prosody," *Speech Commun.*, vol. 33, pp. 135–151, 2001.

[13] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: Studies on speech corpora," in *Proc. AAAI Spring Symp. Empirical Methods in Discourse Interpretation and Generation*, Mar. 1995, pp. 106–112.

[14] G. Ferguson, J. Allen, B. Miller, and E. Ringger, "The design and implementation of the TRAINS-96 system: A prototype mixed-initiative planning assistant," Univ. of Rochester, Rochester, Tech. Rep. TN96-5, Oct. 1996.

[15] M. Ostendorf, I. Shafran, S. Shattuck-Hufnagel, L. Carmichael, and W. Byrne, "A prosodically labeled database of spontaneous speech," in *Proc. ISCA Workshop Prosody in Speech Recognition and Understanding*, Oct. 2001, pp. 119–121.

[16] S. Ananthakrishnan and S. Narayanan, "A novel algorithm for unsupervised prosodic language model adaptation," in *Proc. Int. Conf. Acoust. , Speech Signal Process.*, Las Vegas, NV, 2008.

[17] "CSR-II (WSJ1) Complete," Linguistic Data Consortium, Philadelphia, PA, 1994.

[18] B. Pellom, "SONIC: The University of colorado continuous speech recognizer," Univ. of Colorado, Boulder, CO, Tech. Rep. TR-CSLR-2001-01, Mar. 2001.

[19] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, 2002, vol. 2, pp. 901–904.

[20] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech, Lang.*, vol. 14, no. 4, pp. 373–400, 2000.

[21] G. Evermann, "Minimum word error rate decoding," M.S. thesis, Cambridge Univ., Cambridge, U.K., 1999.

[22] J. Bilmes, "A Gentle tutorial on the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," Univ. of Berkeley, Berkeley, CA, Tech. Rep. ICSI-TR-97-021, 1997.

[23] J.-L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[24] J.-L. Gauvain and C.-H. Lee, "Bayesian learning of Gaussian mixture densities for hidden Markov models," in *Proc. DARPA Speech and Natural Language Workshop*, Pacific Grove, CA, 1991, pp. 272–277, Morgan–Kaufmann.

[25] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland*, The HTK Book*. Cambridge, U.K.: Cambridge Univ., Dec. 2002.

Sankaranarayanan Ananthakrishnan received the B.Eng. degree in electronic and telecommunication engineering from the University of Mumbai, India, in 2002 and the M.Sc. and Ph.D. degrees, both in electrical engineering, from the University of Southern California (USC), Los Angeles, in 2004 and 2008, respectively.

He is currently a Research Scientist in the Speech and Language Technologies Business Unit, BBN Technologies, Cambridge, MA. An alumnus of the Speech Analysis and Interpretation Laboratory (SAIL) at USC,his research interests are broadly centered around spoken language processing and understanding, with a focus on higher level aspects of language, such as prosody. His other interests include speech recognition, natural language processing, speech-to-speech translation, and related areas in statistical pattern recognition and machine learning. He has published over 12 papers in peer-reviewed conferences and journals.

Dr. Ananthakrishnan received the Best Student Paper award at ICASSP 2005.

Shrikanth S. Narayanan (S'88–M'95–SM'02) received the Ph.D. degree from the University of California, Los Angeles, in 1995.

He is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member of its Technical Staff from 1995–2000. At USC, he is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center. He has published over 260 papers and has 14 granted/pending U.S. patents

Dr. Narayanan is a recipient of an NSF CAREER Award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 Best Paper Award from the IEEE Signal Processing Society. Papers by his students have won best student paper awards at ICSLP'02, ICASSP'05, MMSP'06, and MMSP'07. He is an Editor for the *Computer Speech and Language Journal* (2007-present) and an Associate Editor for the *IEEE Signal Processing Magazine*. He was also an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (2000–2004. He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication Committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu.