Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework

Vivek Rangarajan, Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory
University of Southern California
Viterbi School of Electrical Engineering
vrangara@usc.edu,shri@sipi.usc.edu

Srinivas Bangalore
AT&T Research Labs
180 Park Avenue
Florham Park, NJ 07932, U.S.A.
srini@research.att.com

Abstract

In this paper we describe an automatic prosody labeling framework that exploits both language and speech information. We model the syntactic-prosodic information with a maximum entropy model that achieves an accuracy of 85.2% and 91.5% for pitch accent and boundary tone labeling on the Boston University Radio News corpus. We model the acoustic-prosodic stream with two different models, one a maximum entropy model and the other a traditional HMM. We finally couple the syntactic-prosodic and acoustic-prosodic components to achieve significantly improved pitch accent and boundary tone classification accuracies of 86.0% and 93.1% respectively. Similar experimental results are also reported on Boston Directions corpus.

1 Introduction

Prosody refers to intonation, rhythm and lexical stress patterns of spoken language that convey linguistic and paralinguistic information such as emphasis, intent, attitude and emotion of a speaker. Prosodic information associated with a unit of speech, say, syllable, word, phrase or clause, influence all the segments of the unit in an utterance. In this sense they are also referred to as suprasegmentals (Lehiste, 1970). Prosody in general is highly dependent on individual speaker style, gender, dialect and other phonological factors. The difficulty in reliably characterizing suprasegmental information present in speech has resulted in symbolic and parameteric prosody labeling standards like ToBI (Tones and Break Indices) (Silverman et al., 1992) and Tilt model (Taylor, 1998) respectively.

Prosody in spoken language can be characterized through acoustic features or lexical features or both. Acoustic correlates of duration, intensity and pitch, like syllable nuclei duration, short time energy and fundamental frequency (f0) are some acoustic features that are perceived to confer prosodic prominence or stress in English. Lexical features like parts-of-speech, syllable nuclei identity, syllable stress of neighboring words have also demonstrated high degree of discriminatory evidence in prosody detection tasks.

The interplay between acoustic and lexical features in characterizing prosodic events has been successfully exploited in text-to-speech synthesis (Bulyko and Ostendorf, 2001; Ma et al., 2003), speech recognition (Hasegawa-Johnson et al., 2005) and speech understanding (Wightman and Ostendorf, 1994). Text-to-speech synthesis relies on lexical features derived predominantly from the input text to synthesize natural sounding speech with appropriate prosody. In contrast, output of a typical automatic speech recognition (ASR) system is noisy and hence, the acoustic features are more useful in predicting prosody than the hypothesized lexical transcript which may be erroneous. Speech understanding systems model both the lexical and acoustic features at the output of an ASR to improve natural language understanding. Another source of renewed interest has come from spoken language translation (Nöth et al., 2000; Agüero et al., 2006). A prerequisite for all these applications is accurate prosody detection, the topic of the present work.

In this paper, we describe our framework for building an automatic prosody labeler for English. We report results on the Boston University (BU) Radio Speech Corpus (Ostendorf et al., 1995) and Boston Directions Corpus (BDC) (Hirschberg and Nakatani, 1996), two publicly available speech corpora with manual ToBI annotations intended for experiments in automatic prosody labeling. We condition prosody not only on word strings and their parts-of-speech but also on richer syntactic information encapsulated in the form of Supertags (Bangalore and Joshi, 1999). We propose a maximum entropy modeling framework for the syntactic features. We model the acoustic-prosodic stream with two different models, a maximum entropy model and a more traditional hidden markov model (HMM). In an automatic prosody labeling task, one is essentially try-

ing to predict the correct prosody label sequence for a given utterance and a maximum entropy model offers an elegant solution to this learning problem. The framework is also robust in the selection of discriminative features for the classification problem. So, given a word sequence $W = \{w_1, \cdots, w_n\}$ and a set of acoustic-prosodic features $A = \{o_1, \cdots, o_T\}$, the best prosodic label sequence $L^* = \{l_1, l_2, \cdots, l_n\}$ is obtained as follows,

$$L^* = \arg\max_L P(L|A, W) \quad (1)$$

$$= \arg\max_L P(L|W).P(A|L, W) \quad (2)$$

$$\approx \arg\max_L P(L|\Phi(W)).P(A|L, W) \quad (3)$$

where $\Phi(W)$ is the syntactic feature encoding of the word sequence W. The first term in Equation (3) corresponds to the probability obtained through our maximum entropy syntactic model. The second term in Equation (3), computed by an HMM corresponds to the probability of the acoustic data stream which is assumed to be dependent only on the prosodic label sequence.

The paper is organized as follows. In section 2 we describe related work in automatic prosody labeling followed by a description of the data used in our experiments in section 3. We present prosody prediction results from off-the-shelf synthesizers in section 4. Section 5 details our proposed maximum entropy syntactic-prosodic model for prosody labeling. In section 6, we describe our acoustic-prosodic model and discuss our results in section 7. We finally conclude in section 8 with directions for future work.

2 Related work

Automatic prosody labeling has been an active research topic for over a decade. Wightman and Ostendorf (Wightman and Ostendorf, 1994) developed a decision-tree algorithm for labeling prosodic patterns. The algorithm detected phrasal prominence and boundary tones at the syllable level. Bulyko and Ostendorf (Bulyko and Ostendorf, 2001) used a prosody prediction module to synthesize natural speech with appropriate prosody. Verbmobil (Nöth et al., 2000) incorporated prosodic labeling into a translation framework for improved linguistic analysis and speech understanding.

Prosody has typically been represented either symbolically, e.g., ToBI (Silverman et al., 1992) or parametrically, e.g., Tilt Intonation Model (Taylor, 1998). Parametric approaches either restrict the variants of prosody by definition or automatically learn prosodic patterns from data (Agüero et al., 2006). The BU corpus is a widely used corpus with symbolic representation of prosody. The hand-labeled ToBI annotations make this an attractive corpus to perform prosody labeling experiments. The main drawback of this corpus is that it comprises only read speech. Prosody labeling on spontaneous speech corpora like Boston Directions corpus (BDC), Switchboard (SWBD) has garnered attention in (Hirschberg and Nakatani, 1996; Gregory and Altun, 2004).

Automatic prosody labeling has been achieved through various machine learning techniques, such as decision trees (Hirschberg, 1993; Wightman and Ostendorf, 1994; Ma et al., 2003), rule-based systems (Shimei and McKeown, 1999), bagging and boosting on CART (Sun, 2002), hidden markov models (Conkie et al., 1999), neural networks (Hasegawa-Johnson et al., 2005),maximum-entropy models (Brenier et al., 2005) and conditional random fields (Gregory and Altun, 2004).

Prosody labeling of the BU corpus has been reported in many studies (Hirschberg, 1993; Hasegawa-Johnson et al., 2005; Ananthakrishnan and Narayanan, 2005). Hirschberg (Hirschberg, 1993) used a decision-tree based system that achieved 82.4% speaker dependent accent labeling accuracy at the word level on the BU corpus using lexical features. (Ross and Ostendorf, 1996) also used an approach similar to (Wightman and Ostendorf, 1994) to predict prosody for a TTS system from lexical features. Pitch accent accuracy at the word-level was reported to be 82.5% and syllable-level accent accuracy was 80.2%. (Hasegawa-Johnson et al., 2005) proposed a neural network based syntactic-prosodic model and a gaussian mixture model based acoustic-prosodic model to predict accent and boundary tones on the BU corpus that achieved 84.2% accuracy in accent prediction and 93.0% accuracy in intonational boundary prediction. With syntactic information alone they achieved 82.7% and 90.1% for accent and boundary prediction, respectively. (Ananthakrishnan and Narayanan, 2005) modeled the acoustic-prosodic information using a coupled hidden markov model that modeled the asynchrony between the acoustic streams. The pitch accent and boundary tone detection accuracy at the syllable level were 75% and 88% respectively. Our proposed maximum entropy syntactic model outperforms previous work. On the BU corpus, with syntactic information alone we achieve pitch accent and boundary tone accuracy of 85.2% and 91.5% on the same training and test sets used in (Chen et al., 2004; Hasegawa-Johnson et al., 2005). Further, the coupled model with both acoustic and syntactic information results in accuracies of 86.0% and 93.1% respectively. On the BDC corpus, we achieve pitch accent and boundary tone accuracies of 79.8% and 90.3%.

3 Data

The BU corpus consists of broadcast news stories including original radio broadcasts and laboratory sim-

Corpus statistics	BU				BDC			
	f2b	**f1a**	**m1b**	**m2b**	**h1**	**h2**	**h3**	**h4**
# Utterances	165	69	72	51	10	9	9	9
# words (w/o punc)	12608	3681	5058	3608	2234	4127	1456	3008
# pitch accents	6874	2099	2706	2016	1006	1573	678	1333
# boundary tones (w IP)	3916	1059	1282	1023	498	727	361	333
# boundary tones (w/o IP)	2793	684	771	652	308	428	245	216

Table 1: BU and BDC dataset used in experiments

ulations recorded from seven FM radio announcers. The corpus is annotated with orthographic transcription, automatically generated and hand-corrected part-of-speech tags and automatic phone alignments. A subset of the corpus is also hand annotated with ToBI labels. In particular, the experiments in this paper are carried out on 4 speakers similar to (Chen et al., 2004), 2 male and 2 female referred to hereafter as **m1b**, **m2b**, **f1a** and **f2b**. The BDC corpus is made up of elicited monologues produced by subjects who were instructed to perform a series of direction-giving tasks. Both spontaneous and read versions of the speech are available for four speakers **h1**, **h2**, **h3** and **h4** with hand-annotated ToBI labels and automatic phone alignments, similar to the BU corpus. Table 1 shows some of the statistics of the speakers in the BU and BDC corpora.

In Table 1, the pitch accent and boundary tone statistics are obtained by decomposing the ToBI labels into binary classes using the mapping shown in Table 2.

BU Labels	**Intermediate Mapping**	**Coarse Mapping**
H*,!H* L* *,*?,X*?	Single Accent	**accent**
H+!H*,L+H*,L+!H* L*+!H,L*+H	Bitonal Accent	
L-L%,!H-L%,H-L% H-H% L-H% %?,X%?,%H	Final Boundary tone	**btone**
L-,H-,!H- -X?,-?	Intermediate Phrase (IP) boundary	
<,>,no label	**none**	**none**

Table 2: ToBI label mapping used in experiments

In all our prosody labeling experiments we adopt a leave-one-out speaker validation similar to the method in (Hasegawa-Johnson et al., 2005) for the four speakers with data from one speaker for testing and from the other three for training. For the BU corpus, **f2b** speaker was always used in the training set since it contains the most data. In addition to performing experiments on all the utterances in BU corpus, we also perform identical experiments on the train and test sets reported in (Chen et al., 2004) which is referred to as Hasegawa-Johnson et al. set.

4 Baseline Experiments

We present three baseline experiments. One is simply based on chance where the majority class label is predicted. The second is a baseline only for pitch accents derived from the lexical stress obtained through look-up from a pronunciation lexicon labeled with stress. Finally, the third and more concrete baseline is obtained through prosody detection in current speech synthesis systems.

4.1 Prosody labels derived from lexical stress

Pitch accents are usually carried by the stressed syllable in a particular word. Lexicons with phonetic transcription and lexical stress are available in many languages. Hence, one can use these lexical stress markers within the syllables and evaluate the correlation with pitch accents. Eventhough the lexicon has a closed vocabulary, letter-to-sound rules can be derived from it for unseen words. For each word carrying a pitch accent, we find the particular syllable where the pitch accent occurs from the manual annotation. For the same syllable, we predict pitch accent based on the presence or absence of a lexical stress marker in the phonetic transcription. The results are presented in Table 3.

4.2 Prosody labeling with Festival and AT&T Natural Voices® Speech Synthesizer

Festival (Black et al., 1998) and AT&T Natural Voices® (NV) speech synthesizer (att,) are two publicly available speech synthesizers that have a prosody prediction module available. We performed automatic prosody labeling using the two synthesizers to get a baseline.

4.2.1 AT&T Natural Voices® Speech Synthesizer

The AT&T NV® speech synthesizer is a half phone speech synthesizer. The toolkit accepts an input text utterance and predicts appropriate ToBI pitch accent and boundary tones for each of

Corpus	Speaker Set	Prediction Module	Pitch accent		Boundary tone	
			Chance	Accuracy	Chance	Accuracy
BU	Entire Set	**Lexical stress**	54.33	72.64	-	-
		AT&T Natural Voices	54.33	81.51	81.14	89.10
		Festival	54.33	69.55	81.14	89.54
	Hasegawa-Johnson et al. set	**Lexical stress**	56.53	74.10	-	-
		AT&T Natural Voices	56.53	81.73	82.88	89.67
		Festival	56.53	68.65	82.88	90.21
BDC	Entire Set	**Lexical stress**	57.60	67.42	-	-
		AT&T Natural Voices	57.60	68.49	88.90	84.90
		Festival	57.60	64.94	88.90	85.17

Table 3: Classification results of pitch accents and boundary tones (in %) using Festival and AT&T NV® synthesizer

the selected units (in this case, a pair of phones) from the database. We reverse mapped the selected half phone units to words, thus obtaining the ToBI labels for each word in the input utterance. The toolkit uses a rule-based procedure to predict the ToBI labels from lexical information. The pitch accent labels predicted by the toolkit are $L_{\mathbf{accent}} \in \{\mathbf{H}*, \mathbf{L}*, \mathbf{none}\}$ and the boundary tones are $L_{\mathbf{btone}} \in \{\mathbf{L}\text{-}\mathbf{L}\%, \mathbf{H}\text{-}\mathbf{H}\%, \mathbf{L}\text{-}\mathbf{H}\%, \mathbf{none}\}$.

4.2.2 Festival Speech Synthesizer

Festival (Black et al., 1998) is an open-source unit selection speech synthesizer. The toolkit includes a CART-based prediction system that can predict ToBI pitch accents and boundary tones for the input text utterance. The pitch accent labels predicted by the toolkit are $L_{\mathbf{accent}} \in \{\mathbf{H}*, \mathbf{L}+\mathbf{H}*, !\mathbf{H}*, \mathbf{none}\}$ and the boundary tones are $L_{\mathbf{btone}} \in \{\mathbf{L}\text{-}\mathbf{L}\%, \mathbf{H}\text{-}\mathbf{H}\%, \mathbf{L}\text{-}\mathbf{H}\%, \mathbf{none}\}$. The prosody labeling results obtained through both the speech synthesis engines are presented in Table 3. The chance column in Table 3 is obtained by predicting the most frequent label in the data set.

In the next sections, we describe our proposed maximum entropy based syntactic model and HMM based acoustic-prosodic model for automatic prosody labeling.

5 Syntactic-prosodic Model

We propose a maximum entropy approach to model the words, syntactic information and the prosodic labels as a sequence. We model the prediction problem as a classification task as follows: given a sequence of words w_i in a sentence $W = \{w_1, \cdots, w_n\}$ and a prosodic label vocabulary ($l_i \in \mathcal{L}$), we need to predict the best prosodic label sequence $L^* = \{l_1, l_2, \cdots, l_n\}$. We approximate the conditional probability to be within a bounded n-gram context. Thus,

$$L^* = \arg\max_L P(L|W,T,S) \quad (4)$$

$$\approx \arg\max_L \prod_i^n p(l_i|w_{i-k}^{i+k}, t_{i-k}^{i+k}, s_{i-k}^{i+k}) \quad (5)$$

where $W = \{w_1, \cdots, w_n\}$ is the word sequence and $T = \{t_1, \cdots, t_n\}$, $S = \{s_1, \cdots, s_n\}$ are the corresponding part-of-speech and additional syntactic information sequences. The variable k controls the context.

The BU corpus is automatically labeled (and hand-corrected) with part-of-speech (POS) tags. The POS inventory is the same as the Penn treebank which includes 47 POS tags: 22 open class categories, 14 closed class categories and 11 punctuation labels. We also automatically tagged the utterances using the AT&T POS tagger. The POS tags were mapped to function and content word categories [1] which was added as a discrete feature. In addition to the POS tags, we also annotate the utterance with Supertags (Bangalore and Joshi, 1999). Supertags encapsulate predicate-argument information in a local structure. They are composed with each other using substitution and adjunction operations of Tree-Adjoining Grammars (TAGs) to derive a dependency analysis of an utterance and its predicate-argument structure. Even though there is a potential to exploit the dependency structure between supertags and prosody labels as demonstrated in (Hirschberg and Rambow, 2001), for this paper we use only the supertag labels.

Finally, we generate one feature vector ($\mathbf{\Phi}$) for each word in the data set (with local contextual features). The best prosodic label sequence is then,

$$L^* = \arg\max_L \prod_i^n P(l_i|\mathbf{\Phi}) \quad (6)$$

To estimate the conditional distribution $P(l_i|\mathbf{\Phi})$ we use the general technique of choosing the maximum entropy (maxent) distribution that estimates the average of each feature over the training data (Berger et al., 1996). This can be written in terms of Gibbs distribution parameterized with weights λ, where V is the size of the prosodic label set. Hence,

$$P(l_i|\mathbf{\Phi}) = \frac{e^{\lambda_{l_i}.\mathbf{\Phi}}}{\sum_{l=1}^{V} e^{\lambda_{l_i}.\mathbf{\Phi}}} \quad (7)$$

[1] function and content word features were obtained through a look-up table based on POS

Corpus	Speaker Set	Syntactic features	k=3	
			accent	**btone**
BU	Entire Set	correct POS tags	84.75	91.39
		AT&T POS + supertags	84.59	91.34
		Joint Model (w AT&T POS + supertags)	84.60	91.36
	Hasegawa-Johnson et al. set	correct POS tags	85.22	91.33
		AT&T POS + supertags	84.95	91.21
		Joint Model (w AT&T POS + supertags)	84.78	91.54
BDC	Entire Set	AT&T POS + supertags	79.81	90.28
		Joint Model (w AT&T POS + supertags)	79.57	89.76

Table 4: Classification results (%) of pitch accents and boundary tones for different syntactic representation ($k = 3$)

We use the machine learning toolkit LLAMA (Haffner, 2006) to estimate the conditional distribution using maxent. LLAMA encodes multiclass maxent as binary maxent to increase the training speed and to scale the method to large data sets. Each of the V classes in the label set $\mathcal{L}$ is encoded as a bit vector such that, in the vector for class i, the i^{th} bit is one and all other bits are zero. Finally, V one-versus-other binary classifiers are used as follows.

$$P(y|\mathbf{\Phi}) = 1 - P(\bar{y}|\mathbf{\Phi}) = \frac{e^{\lambda_y \cdot \mathbf{\Phi}}}{e^{\lambda_y \cdot \mathbf{\Phi}} + e^{\lambda_{\bar{y}} \cdot \mathbf{\Phi}}} \quad (8)$$

where $\lambda_{\bar{y}}$ is the parameter vector for the anti-label $\bar{y}$. To compute $P(l_i|\mathbf{\Phi})$, we use the class independence assumption and require that $y_i = 1$ and for all $j \neq i, y_j = 0$.

$$P(l_i|\mathbf{\Phi}) = P(y_i|\mathbf{\Phi}) \prod_{j \neq i}^{V} P(y_j|\mathbf{\Phi}) \quad (9)$$

5.1 Joint Modeling of Accents and Boundary Tones

Prosodic prominence and phrasing can also be viewed as joint events occurring simultaneously. Previous work by (Wightman and Ostendorf, 1994) suggests that a joint labeling approach may be more beneficial in prosody labeling. In this scenario, we treat each word to have one of the four labels $l_i \in \mathcal{L} = \{$**accent-btone, accent-none, none-btone, none-none**$\}$. We trained the classifier on the joint labels and then computed the error rates for individual classes. The results of prosody prediction using the set of syntactic-prosodic features for $k = 3$ is shown in Table 4. The joint modeling approach provides a marginal improvement in the boundary tone prediction but is slightly worse for pitch accent prediction.

5.2 Supertagger performance on Intermediate Phrase boundaries

Perceptual experiments have indicated that inter-annotator agreement for ToBI intermediate phrase boundaries is very low compared to full-intonational boundaries (Syrdal and McGory, 2000). Intermediate phrasing is important in TTS applications to synthesize appropriate short pauses to make the utterance sound natural. The significance of syntactic features in the boundary tone prediction prompted us to examine the effect of predicting intermediate phrase boundaries in isolation. It is intuitive to expect supertags to perform well in this task as they essentially form a local dependency analysis on an utterance and provide an encoding of the syntactic phrasal information. We performed this task as a three way classification where $l_i \in \mathcal{L} = \{$**btone, ip, none**$\}$. The results of the classifier on IPs is shown in Table 5.

Model	Syntactic features	IP accuracy
k=2 (bigram context)	correct POS tags	83.25
	AT&T POS tags	83.32
	supertags	83.37
k=3 (trigram context)	correct POS tags	83.30
	AT&T POS tags	83.46
	supertags	83.74

Table 5: Accuracy (in %) obtained by leave-one out speaker validation using IPs as a separate class on entire speaker set

6 Acoustic-prosodic model

We propose two approaches to modeling the acoustic-prosodic features for prosody prediction. First, we propose a maximum entropy framework similar to the syntactic model where we quantize the acoustic features and model them as discrete sequences. Second, we use a more traditional approach where we train continuous observation density HMMs to represent pitch accents and boundary tones. We first describe the features used in the acoustic modeling followed by a more detailed description of the acoustic-prosodic model.

6.1 Acoustic-prosodic features

The BU corpus contains the corresponding acoustic-prosodic feature file for each utterance. The f0, RMS energy (e) of the utterance along with features for

Corpus	Speaker Set	Model	Pitch accent		Boundary tone	
			Acoustics	Acoustics+syntax	Acoustics	Acoustics+syntax
BU	Entire Set	Maxent acoustic model	80.09	84.53	84.10	91.56
		HMM acoustic model	70.58	85.13	71.28	92.91
	Hasegawa-Johnson et al. set	Maxent acoustic model	80.12	84.84	82.70	91.76
		HMM acoustic model	71.42	86.01	73.43	93.09
BDC	Entire Set	Maxent acoustic model	74.51	78.64	83.53	90.49

Table 6: Classification results of pitch accents and boundary tones (in %) with acoustics only and acoustics+syntax using both our models

distinction between voiced/unvoiced segment, cross-correlation values at estimated f0 value and ratio of first two cross correlation values are computed over 10 msec frame intervals. In our experiments, we use these values rather than computing them explicitly which is straightforward with most audio toolkits. Both the energy and the f0 levels were normalized with speaker specific means and variances. Delta and acceleration coefficients were also computed for each frame. The final feature vector is 6-dimensional comprising of f0, Δf0, Δ^2f0, e, Δe, Δ^2e per frame.

6.2 Maximum Entropy acoustic-prosodic model

We propose a maximum entropy modeling framework to model the continuous acoustic-prosodic observation sequence as a discrete sequence through the means of quantization. The quantized acoustic stream is then used as a feature vector and the conditional probabilities are approximated by an n-gram model. This is equivalent to reducing the vocabulary of the acoustic-prosodic features and hence offers better estimates of the conditional probabilities. Such an n-gram model of quantized continuous features is similar to representing the set of features with a linear fit as done in the tilt intonational model (Taylor, 1998).

The quantized acoustic-prosodic feature stream is modeled with a maxent acoustic-prosodic model similar to the one described in section 5. Finally, we append the syntactic and acoustic features to model the combined stream with the maxent acoustic-syntactic model, where the objective criterion for maximization is Equation (1). The pitch accent and boundary tone prediction accuracies for quantization performed by considering only the first decimal place is reported in Table 6. As expected, we found the classification accuracy to drop with increasing number of bins used in the quantization due to the small amount of training data.

6.3 HMM acoustic-prosodic model

We also investigated the traditional HMM approach to model the high variability exhibited by the acoustic-prosodic features. First, we trained separate context independent single state Gaussian mixture density HMMs for pitch accents and boundary tones in a generative framework. The label sequence was decoded using the viterbi algorithm. Next, we trained HMMs with 3 state left-to-right topology with uniform segmentation. The segmentations need to be uniform due to lack of an acoustic-prosodic model trained on the features pertinent to our task to obtain forced segmentation.

The final label sequence using the maximum entropy syntactic-prosodic model and the HMM based acoustic-prosodic model was obtained by combining the syntactic and acoustic probabilities shown in Equation (3). The syntactic-prosodic maxent model outputs a posterior probability for each class per word. We formed a lattice out of this structure and composed it with the lattice generated by the HMM acoustic-prosodic model. The best path was chosen from the composed lattice through a Viterbi search. The acoustic-prosodic probability $P(A|L,W)$ was raised by a power of γ to adjust the weighting between the acoustic and syntactic model. The value of γ was chosen as 0.008 and 0.015 for pitch accent and boundary tone respectively, by tuning on the training set. The results of the acoustic-prosodic model and the coupled model are shown in Table 6.

7 Discussion

The baseline experiment with lexical stress obtained from a pronunciation lexicon for prediction of pitch accent yields substantially higher accuracy than chance. This could be particularly useful in resource-limited languages where prosody labels are usually not available but one has access to a reasonable lexicon with lexical stress markers. Off-the-shelf speech synthesizers like Festival and AT&T speech synthesizer perform reasonably well in pitch accent and boundary tone prediction. AT&T speech synthesizer performs better than Festival in pitch accent prediction and the latter performs better in boundary tone prediction. This can be attributed to better rules in the AT&T synthesizer for pitch accent prediction. Boundary tones are usually highly correlated with punctuation and Festival seems to capture this well. However, both these synthesizers generate a high de-

gree of false alarms.

Our syntactic-prosodic maximum entropy model proposed in section 5 outperforms previously reported results on pitch accent and boundary tone classification. Much of the gain comes from the robustness of the maximum entropy modeling in capturing the uncertainty in the classification task. Considering the inter-annotator agreement for ToBI labels is only about 81% for pitch accents and 93% for boundary tones, the maximum entropy framework is able to capture the uncertainty present in manual annotation. The supertag feature offers additional discriminative information over the part-of-speech tags (also as shown by (Hirschberg and Rambow, 2001).

The maximum entropy acoustic-prosodic model discussed in section 6.2 performs reasonably well in isolation. This is a simple method and the quantization resolution can be adjusted based on the amount of data available for training. However, the model does not perform as well when combined with the syntactic features. We conjecture that the generalization provided by the acoustic HMM model is complementary to that provided by the maximum entropy model, resulting in better accuracy when combined together as compared to that of a maxent-based acoustic and syntactic model.

The weighted maximum entropy syntactic-prosodic model and HMM acoustic-prosodic model performs the best in pitch accent and boundary tone classification. The classification accuracies are as good as the inter-annotator agreement for the ToBI labels. Our HMM acoustic-prosodic model is a generative model and does not assume the knowledge of word boundaries in predicting the prosodic labels as in most approaches (Hirschberg, 1993; Wightman and Ostendorf, 1994; Hasegawa-Johnson et al., 2005). This makes it possible to have true parallel prosody prediction during speech recognition. The weighted approach also offers flexibility in prosody labeling for either speech synthesis or speech recognition. While the syntactic-prosodic model would be more discriminative for speech synthesis, the acoustic-prosodic model is more appropriate for speech recognition.

8 Conclusions and Future Work

In this paper, we described a maximum entropy modeling framework for automatic prosody labeling. We presented two schemes for prosody labeling that utilize the acoustic and syntactic information from the input utterance, a maximum entropy model that models the acoustic-syntactic information as a sequence and the other that combines the maximum entropy syntactic-prosodic model and a HMM based acoustic-prosodic model. We also used enriched syntactic information in the form of supertags in addition to POS tags. The supertags provide an improvement in both the pitch accent and boundary tone classification. Especially, in the case where the input utterance is automatically POS tagged (and not hand-corrected), supertags provide a marginal but definite improvement in prosody labeling. The maximum entropy syntactic-prosodic model alone resulted in pitch accent and boundary tone accuracies of 85.2% and 91.5% on training and test sets identical to (Chen et al., 2004). As far as we know, these are the best results on the BU corpus using syntactic information alone and a train-test split that does not contain the same speakers. The acoustic-syntactic maximum entropy model performs better than its syntactic-prosodic counterpart for the boundary tone case but is slightly worse for pitch accent scenario partly due to the approximation involved in quantization. But these results are still better than the baseline results from out-of-the-box speech synthesizers. Finally, our combined maximum entropy syntactic-prosodic model and HMM acoustic-prosodic model performs the best with pitch accent and boundary tone labeling accuracies of 86.0% and 93.1% respectively.

As a continuation of our work, we are incorporating our automatic prosody labeler in a speech-to-speech translation framework. Typically, state-of-the-art speech translation systems have a source language recognizer followed by a machine translation system. The translated text is then synthesized in the target language with prosody predicted from text. In this process, some of the critical prosodic information present in the source data is lost during translation. With reliable prosody labeling in the source language, one can transfer the prosody to the target language (this is feasible for languages with phrase level correspondence). The prosody labels by themselves may or may not improve the translation accuracy but they provide a framework where one can obtain prosody labels in the target language from the speech signal rather than depending on a lexical prosody prediction module in the target language.

Acknowledgements

We would like to thank Vincent Goffin, Stephan Kanthak, Patrick Haffner, Enrico Bocchieri for their support with acoustic modeling tools. We are also thankful to Alistair Conkie, Yeon-Jun Kim, Ann Syrdal and Julia Hirschberg for their help and guidance with the synthesis components and ToBI labeling standard.

References

P. D. Agüero, J. Adell, and A. Bonafonte. 2006. Prosody generation for speech-to-speech transla-

tion. In *Proceedings of ICASSP*, Toulouse, France, May.

S. Ananthakrishnan and S. Narayanan. 2005. An automatic prosody recognizer using a coupled multistream acoustic model and a syntactic-prosodic language model. In *In Proceedings of ICASSP*, Philadelphia, PA, March.

AT&T Natural Voices speech synthesizer. http://www.naturalvoices.att.com.

S. Bangalore and A. K. Joshi. 1999. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2), June.

A. Berger, S. D. Pietra, and V. D. Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

A. W. Black, P. Taylor, and R. Caley. 1998. The Festival speech synthesis system. http://festvox.org/festival.

J. M. Brenier, D. Cer, and D. Jurafsky. 2005. The detection of emphatic words using acoustic and lexical features. In *In Proceedings of Eurospeech*.

I. Bulyko and M. Ostendorf. 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In *Proc. of ICASSP*.

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

A. Conkie, G. Riccardi, and R. C. Rose. 1999. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In *Proc. Eurospeech*, pages 523–526, Budapest, Hungary.

M. Gregory and Y. Altun. 2004. Using conditional random fields to predict pitch accent in conversational speech. In *42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

P. Haffner. 2006. Scaling large margin classifiers for spoken language understanding. *Speech Communication*, 48(iv):239–261.

M. Hasegawa-Johnson, K. Chen, J. Cole, S. Borys, S. Kim, A. Cohen, T. Zhang, J. Choi, H. Kim, T. Yoon, and S. Chavara. 2005. Simultaneous recognition of words and prosody in the boston university radio speech corpus. *Speech Communication*, 46:418–439.

J. Hirschberg and C. Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Proceedings of the 34th conference on Association for Computational Linguistics*, pages 286–293.

J. Hirschberg and O. Rambow. 2001. Learning prosodic features using a tree representation. In *Proceedings of Eurospeech*, pages 1175–1180, Aalborg.

J. Hirschberg. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2).

I. Lehiste. 1970. *Suprasegmentals*. MIT Press, Cambridge, MA.

X. Ma, W. Zhang, Q. Shi, W. Zhu, and L. Shen. 2003. Automatic prosody labeling using both text and acoustic information. In *Proceedings of ICASSP*, volume 1, pages 516–519, April.

E. Nöth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann. 2000. VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system. *IEEE Transactions on Speech and Audio processing*, 8(5):519–532.

M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. 1995. The Boston University Radio News Corpus. Technical Report ECS-95-001, Boston University, March.

K. Ross and M. Ostendorf. 1996. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, Oct.

P. Shimei and K. McKeown. 1999. Word informativeness and automatic pitch accent modeling. In *In Proceedings of EMNLP/VLC*, College Park, Maryland.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *Proceedings of ICSLP*, pages 867–870.

X. Sun. 2002. Pitch accent prediction using ensemble machine learning. In *Proc. of ICSLP*.

A. K. Syrdal and J. McGory. 2000. Inter-transcriber reliability of tobi prosodic labeling. In *Proc. ICSLP*, pages 235–238, Beijing, China.

P. Taylor. 1998. The tilt intonation model. In *Proc. ICSLP*, volume 4, pages 1383–1386.

C. W. Wightman and M. Ostendorf. 1994. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2(3):469–481.