ACOUSTIC-SYNTACTIC MAXIMUM ENTROPY MODEL FOR AUTOMATIC PROSODY LABELING

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 intended for the purpose of incorporating prosody within a speech-to-speech translation framework. We propose a maximum entropy syntactic-prosodic model that achieves an accuracy of 85.22% and 91.54% 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 a pitch accent and boundary tone classification accuracy of 86.01% and 93.09% respectively.

1. INTRODUCTION

Prosody refers to the rhythm and intonation patterns of spoken language that convey meaningful information beyond the orthographic transcription. In this sense they are also referred to as suprasegmentals [1], that convey both linguistic and paralinguistic information like emphasis, intent, attitude and emotion of a speaker. Acoustic correlates of duration, intensity and pitch like syllable duration, short time energy and fundamental frequency (f0) are perceived to confer prosodic prominence or stress in English. However, these prosodic cues cannot be quantified in an absolute manner and are highly relative to individual speaker style, gender, dialect and other phonological factors. The difficulty in reliably characterizing suprasegmental information present in speech signal has resulted in prosodic labeling standards like ToBI for American English [2].

Automatic recognition and identification of prosodic events is vital in text-to-speech (TTS) synthesis [3], speech understanding [4], speech recognition [5] and speech-to-speech translation [6, 7] applications. While automatic prosody labeling has been actively pursued over the last several years (see Sec. 2), one source of renewed interest has come from recent spoken language translation applications. The work described in this paper is motivated by the desire to incorporate prosody within a speech-to-speech translation framework. Typically, state-of-the-art speech translation systems have a source language recognizer followed by a translator. The translated text is then synthesized in the target language with prosody predicted from text. In this process, the prosodic information present in the source signal is lost during translation. However, with reliable prosody labeling in the source language, the prosody can be transferred to the target language (e.g., English-to-Spanish) and the predicted prosody can used by a TTS system in synthesizing speech with appropriate prosody. A pre-requisite for such applications is the accurate prosody labeling, the topic of the present work.

In this paper, we describe the first phase of our work that entails building an automatic prosody labeler for the source language (English in our case). We use the Boston University (BU) Radio Speech Corpus [8], one of several 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 [9]. 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 trying 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) corresponds to the probability of the acoustic data stream which is assumed to be dependent only on the prosodic label sequence obtained through a HMM.

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 conclude in section 7 with directions for future work.

2. RELATED WORK

Automatic prosody labeling has been an active research topic for over a decade. Wightman and Ostendorf [4] developed a decision-

tree algorithm for labeling prosodic patterns. The algorithm detected phrasal prominence and boundary tones at the syllable level. Bulyko and Ostendorf [3] used a prosody prediction module to synthesize natural speech with appropriate prosody. Verbmobil [6] incorporated prosodic labeling into a translation framework for improved linguistic analysis and speech understanding.

Automatic prosody labeling within the BU corpus has been reported in many studies [5, 10, 11]. Hirschberg [10] 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 [12] also used an approach similar to [4] 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., [5] 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 84.21% accuracy in accent prediction and 93.07% accuracy in intonational boundary prediction. With syntactic information alone they achieved 82.67% and 90.09% for accent and boundary prediction, respectively. Ananthakrishnan and Narayanan, [11] modeled the acoustic-prosodic information using a coupled hidden markov model that models 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. With syntactic information alone we achieve pitch accent and boundary tone accuracy of 85.22% and 91.54% on the same training and test sets used in [13]. Further, the coupled model with both acoustic and syntactic information results in accuracies of 86.01% and 93.09% respectively.

3. DATA

The BU corpus consists of broadcast news stories including original radio broadcasts and laboratory simulations 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 [13], 2 male and 2 female referred to hereon as **m1b**, **m2b**, **f1a** and **f2b**. The following table shows some of the statistics of the speakers in the BU corpus:

Speakers	**f2b**	**f1a**	**m1b**	**m2b**
# Utterances	165	69	72	51
# words (w/o punc)	12608	3681	5058	3608
# pitch accents	6874	2099	2706	2016
# boundary tones (w IP)	3916	1059	1282	1023
# boundary tones (w/o IP)	2793	684	771	652

Table 1: Boston University dataset used in experiments

In Table 1, the pitch accent and boundary tone statistics are obtained by decomposing the 26 BU dataset ToBI labels into binary classes using the mapping shown in Table 2. In all our prosody labeling experiments we adopt a leave-one-out speaker validation similar to the method in [5] for the four speakers with data from one speaker for testing and from the other three for training. Also, **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 [13] which is referred to as Hasegawa-Johnson et al. set.

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: Boston University label mapping used in experiments

4. PROSODY LABELING WITH FESTIVAL AND AT&T NATURAL VOICES® SPEECH SYNTHESIZER

Festival [14] and AT&T Natural Voices® (NV) speech synthesizer [15] 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.

4.1. AT&T Natural Voices® Speech Synthesizer

The AT&T NV® speech synthesizer is a diphone-based speech synthesizer. The toolkit accepts an input text utterance and predicts appropriate ToBI pitch accent and boundary tones for each of the selected units (in this case, a pair of phones) from the database. We reverse mapped the selected diphone 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 $L_{\textbf{accent}} \in \{\mathbf{H*}, \mathbf{L*}, \mathbf{none}\}$ and the boundary tones are $L_{\textbf{btone}} \in \{\mathbf{L\text{-}L\%}, \mathbf{H\text{-}H\%}, \mathbf{L\text{-}H\%}, \mathbf{none}\}$.

4.2. Festival Speech Synthesizer

Festival [14] 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 $L_{\textbf{accent}} \in \{\mathbf{H*}, \mathbf{L+H*}, \mathbf{!H*}, \mathbf{none}\}$ and the boundary tones are $L_{\textbf{btone}} \in \{\mathbf{L\text{-}L\%}, \mathbf{H\text{-}H\%}, \mathbf{L\text{-}H\%}, \mathbf{none}\}$. The prosody labeling results are presented in Table 3. The baseline column in Table 3 is obtained by predicting the most frequent label in the data set.

		Pitch accent		**Boundary tone**	
Speaker Set	Prediction Module	Baseline	Accuracy	Baseline	Accuracy
Entire Set	**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	**AT&T Natural Voices**	56.53	81.73	82.88	89.67
	Festival	56.53	68.65	82.88	90.21

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

In the next section, 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 ($l_i \in L$), we need to predict the

best prosodic label sequence $L^* = \{l_1, l_2, \cdots, l_n\}$. We approximate the conditional probability with an n-gram language model to cope with the prediction errors of the classifier. 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 [9]. 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 [16], 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 [17]. This can be written in terms of Gibbs distribution parameterized with weights λ, where V is the size of the prosodic label set. Hence,

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

We use the machine learning toolkit LLAMA [18] 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 Ł is encoded as a bit vector such that, in the vector for class i, the i^{ith} 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 $j \neq i, y_j = 0$.

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

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

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 [4] 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 $l_i \in L = \{$**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.

Speaker Set	Syntactic features	k=3	
		accent	**btone**
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

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

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 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. The final model is 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 5. As expected, we found the classification accuracy to drop with increasing number of bins used in the quantization.

		Pitch accent		Boundary tone	
Speaker Set	Model	Acoustics	Acoustics+syntax	Acoustics	Acoustics+syntax
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

Table 5: Classification results of pitch accents and boundary tones (in %) with acoustics only and acoustics+syntax using both our models

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 5.

7. 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 85.22% and 91.54% on training and test sets identical to [13]. 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 86.01% and 93.09% respectively.

As a continuation of our work, we are incorporating our automatic prosody labeler in a speech-to-speech translation framework. 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.

8. REFERENCES

[1] I.Lehiste, *Suprasegmentals*. Cambridge, MA: MIT Press, 1970.

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

[3] I. Bulyko and M. Ostendorf, "Joint prosody prediction and unit selection for concatenative speech synthesis," in *Proc. of ICASSP*, 2001.

[4] C.W.Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 469–481, 1994.

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

[6] E. Noth, A. Batliner, A. Kießling, R. Kompe, and H. Niemann, "VERBMOBIL: The use of prosody in the linguistic components of a speech understanding system," *IEEE Transactions on Speech and Audio processing*, vol. 8, no. 5, pp. 519–532, 2000.

[7] P. D. Agüero, J. Adell, and A. Bonafonte, "Prosody generation for speech-to-speech translation," in *Proceedings of ICASSP*, (Toulouse, France), May 2006.

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

[9] S. Bangalore and A. Joshi, "Supertagging: An approach to almost parsing," *Computational Linguistics*, vol. 25, June 1999.

[10] J.Hirschberg, "Pitch accent in context: Predicting intonational prominence from text," *Artificial Intelligence*, vol. 63, no. 1-2, 1993.

[11] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *In Proceedings of ICASSP*, (Philadelphia, PA), March 2005.

[12] K.Ross and M.Ostendorf, "Prediction of abstract prosodic labels for speech synthesis," *Computer Speech and Language*, vol. 10, pp. 155–185, Oct. 1996.

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

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

[15] "AT&T Natural Voices speech synthesizer," http://www.naturalvoices.att.com.

[16] J. Hirschberg and O. Rambow, "Learning prosodic features using a tree representation," in *Proceedings of Eurospeech*, pp. 1175–1180, 2001.

[17] A.Berger, S.D.Pietra, and V.D.Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[18] P.Haffner, "Scaling large margin classifiers for spoken language understanding," *Speech Communication*, vol. 48, no. iv, pp. 239–261, 2006.