AN AUTOMATIC PROSODY LABELING SYSTEM USING ANN-BASED SYNTACTIC-PROSODIC MODEL AND GMM-BASED ACOUSTIC-PROSODIC MODEL

Ken Chen, Mark Hasegawa-Johnson and Aaron Cohen

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, U.S.A.
{kenchen, jhasegaw, ascohen}@uiuc.edu

ABSTRACT

Automatic prosody labeling is important for both speech synthesis and automatic speech understanding. Humans use both syntactic cues and acoustic cues to develop their prediction of prosody for a given utterance. This process can be effectively modeled by an ANN-based syntactic-prosodic model that predicts prosody from syntax and a GMM-based acoustic-prosodic model that predicts prosody from acoustic-prosodic observations. Our experiments on the Radio News Corpus show that ANN is effective in learning the stochastic mapping from the syntactic representation of word strings to prosody labels, with an accuracy of 82.7% for pitch accent labeling and 90.5% for intonational phrase boundary (IPB) labeling. When acoustic observations and reasonably accurate phoneme transcriptions are given, a GMM-based acoustic-prosodic model, coupled with the syntactial-prosodic model, can achieve 84% pitch accent recognition accuracy and 93% IPB recognition accuracy. These results are obtained using different speakers for training and testing and have considerably exceeded all previously reported results on the same corpus, especially for the task of IPB detection.

1. INTRODUCTION

Prosody refers to the suprasegmental features of natural speech (such as rhythm and intonation) that are used to convey linguistic and paralinguistic information (such as emphasis, intention, attitude and emotion). Automatic prosody labeling is important for both speech synthesis and automatic speech understanding.

High quality text-to-speech synthesis systems require accurate prosody labels to generate natural-sounding speech. In these systems, prosody is assigned based on information extracted from text. Although it is generally believed that syntactic, semantic and pragmatic factors are all involved in prosody decision, such labeling relies primarily on syntactic analysis due to the difficulty of representing and extracting high-level linguistic information (the discourse, pragmatic, semantical information) from text. Hirschberg [1] has proposed a decision-tree based system that achieved 82.4% speaker dependent accent labeling accuracy on Radio News, a large improvement over early systems that label prosody based on function word versus content word distinction. Hirschberg's result is important because it shows that it is possible to accurately predict prosody from syntax. In another corpus-based study, Arnfield [2] claimed, after his bigram models predicted prosodic stress from parts-of-speech (POS) with 91% accuracy, that although differing prosodies are possible for a fixed syntax, the syntax of an utterance can be used to generate an underlying "baseline" prosody regardless of actual words, semantics or context. Similar results were achieved by Ross [3], whose system predicts ToBI [4] style prosody labels from text with 82.5% word-level accent presence/absence accuracy. Ross's decision-tree based system is different from Hirschberg's in that it assigns prosody at syllable level instead of at word level and requires pre-generated prosodic phrase structure as input. Even though the importance of syntax in predicting prosody has been recognized in designing these previous systems, the syntactic information contained in the text are not fully utilized: these systems either used small POS set (only 8 POS categories in [1] [3]) due to the limitation in their decision-tree algorithm, or included only small POS context (unigram in [1] [3] and bigram in [2]).

In automatic speech understanding, prosody has been widely used to infer the status of the high-level linguistic units such as syntax, disfluency, dialog act, semantics and emotion. Two prosody recognition models are usually desired: 1. a syntactic-prosodic model that recognizes prosody from the recognized word strings, 2. an acoustic-prosodic model that recognizes prosody from the acoustic signal. The syntactic-prosodic model is essentially the same as what has been used in speech synthesis except that the recognized text instead of the original text is used as input. The acoustic-prosodic model models the PDFs of the acoustic-prosodic observation over given words or phonemes and can be used alone or coupled with the syntactic-prosodic model to improve prosody recognition performance. An early example for the acoustic-prosodic modeling was reported by Wightman et al. [5]. In their system, a decision-tree and a Markov chain model were used to compute the probability of syllable-level prosody sequences given the syllable-timed acoustic features. Their system does not have a syntactic-prosodic model and assumed that prosody can be determined completely from the acoustic correlates (the pitch, duration and energy, etc.) and lexical stress information. This system has achieved success in labeling pitch accents on Radio News Corpus with 84% accuracy on accent presence/absence prediction, higher than the chance level 55% (the percentage of unaccented words out of all words). However, it performed poorly on intonational phrase boundary (IPB) recognition: IPB recognition rate is only 71%, below the chance level of 83% (the percentage of IPBs out of all word boundaries). The failure of IPB detection is mainly caused by the insufficiency of acoustic statistics around the IPBs.

The acoustic-prosodic features are known to be highly variable not only in their strength (amplitude, shape, duration) but also in their time alignment with the syllables (e.g., the peak or valley of the pitch contour may occur on the syllables preceding or succeeding the accented syllable). In addition, they often suffer from both

inter-speaker difference (e.g., some speakers use more expressive prosody than others) and intra-speaker difference (e.g., a speaker can use different prosody for the same word strings in different context). In fact, determining prosody labels from only their local acoustic context is not only difficult for machines but also difficult for human labelers. While listening to speech, human labelers often utilizes the extracted syntactic and semantic information to determine the plausible prosodic labels.

Kompe [6] has proposed another prosody recognition system that uses neural network for the acoustic-prosodic modeling of phoneme-wise prosody and a polygram model for the syntactic-prosodic modeling of word-wise prosody. The polygram model that he used computes the probability of a prosody label p_l given the surrounding n words: $p(p_l|w_{l-n+2}, w_{l-n+3}, \ldots, w_{l+n-1})$. Kompe's system achieved 95% IPB recognition rate for his prosodic-syntactic M labels, labels that are deterministically transformed from syntactic clause boundaries (based on a set of empirical rules) but better correlate with prosodic phrase boundaries than syntactic phrase boundaries. Kompe's syntactic-prosodic model would be ideal given a large amount of training data. In practice, conditioning prosody on word strings creates problems of data-sparseness, especially for small-sized corpora. Despite this disadvantage, Kompe's result suggests the potential advantage of modeling the dependence of prosody over large context ($n > 3$) and relatively large variety of word categories rather than the over-simplified POS classes. Rather than conditioning prosody on word strings, conditioning prosody on their syntactic representation (e.g., parts-of-speech) can effectively reduce the entropy of the syntactic-prosodic models [7].

Motivated by these results, we propose to build a syntactic-prosodic model using artificial neural networks (ANN), an acoustic-prosodic model using Gaussian mixture models (GMM), and a model that couples the syntactic and acoustic models together as a maximum likelihood recognizer. In section 2, we discuss these models in details. Section 3 reports the experiments and the results and conclusions are given in section 4.

2. METHOD

2.1. The syntactic-prosodic model

We propose to model the stochastic mapping from syntactic context to prosody using a multilayer perceptron (MLP) where MLP is used to compute the posterior probability of the prosody label:

$$p(p_l = i|\phi_l(W)) = \frac{g_i(\phi_l(W))}{\sum_i g_i(\phi_l(W))}. \quad (1)$$

In equation (1), p_l represents the prosody of the l^{th} word in the word sequence W, $g_i(\cdot)$ is the i^{th} output of the MLP, $\phi_l(W)$ represents the syntactic information contained in the entire word sequence W that potentially affects the prediction of p_l. Heuristically, $\phi_l(W)$ is chosen such that it contains syntactic information from a fixed window of n words surrounding p_l:

$$\phi_l(W) = (s_{l-(n-1)/2}, \ldots, s_l, \ldots, s_{l+(n-1)/2}), \quad (2)$$

where s_l represents the syntactic information contained in w_l that affects the prediction of p_l. In general, s_l can include all possible information one could obtain from the text analysis (including semantic information). Parts-of-speech is shown to be most useful, but other type of information such as the location of syntactic boundaries is also helpful. The number of output nodes is determined by the variety of prosodic distinctions modeled at word level. In this paper, we chose to model only four possible prosody distinctions for each word: unaccented IPB-medial, unaccented IPB-final, accented IPB-medial and accented IPB-final. This set of prosody labels are simplified from original ToBI labels [4] and are the same as those used in Wightman's study [5].

The syntactic representation used in our experiment for each of these n words includes:

1. parts-of-speech,
2. The number of syntactic phrases the word initiates (phrase opening),
3. The number of syntactic phrases the word terminates (phrase closing).

A set of 32 POS tags are used, which are the same as those used in the Penn Treebank. Syntactic phrase structure is automatically labeled by Charniak's syntactic parser [8]. Since "silence" is annotated in our word transcription, we augmented our parts-of-speech set to include a new label "SIL" which is shown to be very useful for boundary prediction. The "pause" and "breath" cues are among those that are most robust for boundary prediction. If they are not annotated in word transcription, they can be inferred from punctuation.

Each POS tag is mapped to a 33 dimensional binary feature vector. The features for the phrase opening and closing are integer-valued and are normalized to real numbers after being divided by a constant. Each MLP input vector hence contains $35 \times n$ syntactic features. This MLP-based syntactic prosodic model is trained using standard error back-propagation algorithm.

2.2. The acoustic-prosodic model

To recognize p_l from acoustic signal, an acoustic-prosodic model $p(y_l|w_l, p_l)$ can be used, where y_l denotes the acoustic-prosodic observations over w_l. Training this word-level prosody model $p(y_l|w_l, p_l)$ requires lots of data. Instead, it can be computed from its component allophonic acoustic-prosodic models:

$$p(y_l|w_l, p_l) = \sum_{Q_l, H_l} \prod_{(q_k, h_k) \in (Q_l, H_l)} p(y_k|q_k, h_k) p(Q_l, H_l|w_l, p_l) \quad (3)$$

where $p(Q_l, H_l|w_l, p_l)$ is a pronunciation model representing the probability of an allophone string $Q_l = (q_1, \ldots, q_k, \ldots, q_{N_l})$ and associated prosody string $H = (h_1, \ldots, h_k, \ldots, h_{N_l})$ given prosody dependent word token (w_l, p_l), and y_k represents the acoustic-prosodic observation over the allophone q_k. All the prosody-dependent pronunciations are pre-compiled in a lexicon. Note that lexical stress information is conveniently expressed in this pronunciation model, as is the prosody dependent pronunciation variation (different pronunciation of a word under different prosody). An example is given below for the word "above":

- above: ax b ah v
- above!: ax b! ah! v!
- aboveB4: ax b ahB4 vB4
- above!B4: ax b! ah!B4 v!B4

In our labeling scheme, a postfix "!" is used to label the pitch accent for both words and phonemes, and a postfix "B4" is used to label the words and phonemes that are affected by the intonational phrase boundaries. "!" is attached to the phonemes in the primary lexically stressed syllable because in most cases, only the primary lexically stressed syllable in an accented word is accented. We also assumed that preboundary lengthening only occurs in the rhyme of the last syllables in the pre-IPB words [9]. Therefore, only the last rhyme of the preboundary words have "B4" attached. Since a prosody dependent word token (w_l, p_l) may have multiple pronunciations, a summation over (Q_l, H_l) is included in equation (3) to sum up all possible lexical entries for (w_l, p_l).

The primary acoustic cues for prosody are pitch, duration and energy. Other acoustic cues such as voice quality are useful in general but are hard to reliably estimate. In our study, the raw f0 and RMS energy values are obtained using Entropic XWAVES, commercial software well-known for its high-accuracy pitch tracker. Duration features are obtained using the time-aligned phoneme transcription either generated by hand or by automatic methods.

It is important to normalize the pitch and energy features such that they are least affected by both inter-speaker and intra-speaker register variation. The raw $f_0(t)$ returned by pitch tracker is usually noisy and contains pitch doubling and halving errors. To remove these errors, we trained a 3 mixture Gaussian classifier with component means restricted to be $1/2$, 1, and 2 times the utterance mean $\bar{f}_0$. The $f_0(t)$ values classified as samples from either the doubling cluster or halving cluster are removed. $f_0(t)$ are then divided by $\bar{f}_0$ and converted to log scale:

$$\hat{f}_0(t) = \log(f_0(t)/\bar{f}_0 + 1). \tag{4}$$

The $\hat{f}_0(t)$ values that have small probability of voicing (PV), are normally extracted from the non-vocalic frames and are not reliable. Thus, we eliminated the $\hat{f}_0(t)$ whose PVs are smaller than an empirical threshold. We then linearly interpolated $\hat{f}_0(t)$ to recover the complete f_0 contour where the original measures has been previously removed. The linear interpolation of f_0 was proposed by Kompe [6] and has been shown to be a good normalization method. Frame-wise RMS energy values are normalized similarly.

$\hat{f}_0(t)$ is further normalized by an MLP-based nonlinear transformation function $\psi(\cdot)$ trained to minimize the mean square error between the transformed feature $\tilde{f}_0(t)$ and a teaching signal that indicates the location of the transcribed pitch accents. Our experiments suggest that this nonlinear transformation:

$$\tilde{f}_0(t) = \psi(\hat{f}_0(t)), \tag{5}$$

has considerably reduced the intra-speaker differences, especially the pitch declination effects (the gradual reduction of mean and variance of f_0 toward the end of a prosodic phrase) which is known to hurt the accent prediction.

Finally, we computed a group of five features as our base feature set, measured once per segment:

1. allophone duration,
2. average allophone duration over a window of 3 phones,
3. average energy over a window of 3 allophones,
4. the delta of the 3-phone-average of the phoneme-wise mean $\tilde{f}_0$,
5. the delta of item 4.

These features are similar to those in the previous works [6, 9] and are shown to give the best performance among a set of around 15 features. The base feature vectors are rotated using principle component analysis (PCA) such that they can be better modeled by diagonal covariance mixture Gaussians. The deltas of the rotated feature vectors are also included to introduce context dependence. Standard expectation maximization (EM) algorithms are used to train the allophonic acoustic-prosodic models $p(y|q,h)$. Since the duration features are dependent on input phoneme alignment, we tested this system using automatic phoneme transcriptions generated using HMM forced alignment. Only a small degradation in performance (less than 1%) is observed. We believe that this is mainly due to the time-insensitive nature of pitch accents and the already low performance of IPB recognition using only acoustic model. In addition, the large number of average operations in the above feature generating algorithm also helped ameliorate this problem.

2.3. The coupled model

The syntactic-prosodic model and acoustic-prosodic model can be coupled as a maximum likelihood recognizer. Let $W = (w_1, \ldots, w_L)$ be the word sequence, $P = (p_1, \ldots, p_L)$ the prosody sequence of an utterance. The task of prosody recognition is to find the optimal prosody sequence $\tilde{P}$ that maximizes the recognition probability:

$$\begin{aligned}[\tilde{P}] = &\ \arg\max_P p(Y, W), \\ = &\ \arg\max_P p(Y|W, P)p(P|W), \\ = &\ \arg\max_P \prod_{l=1}^{L} p(y_l|w_l, p_l)p(p_l|\phi_l(W))^{\gamma}, \end{aligned} \tag{6}$$

where $Y = (Y_1, \ldots, Y_L)$ is a sequence of L word-wise acoustic-prosodic observation matrices. The syntactic-prosodic probability has been raised by a power of γ, a constant that can be used to adjust the weighting between the syntactic and the acoustic model.

3. EXPERIMENTS AND RESULTS

All our experiments were carried out on the Boston University Radio News Corpus, one of the largest corpora designed for study of prosody [10]. The corpus consists of recordings of broadcast radio news stories including original radio broadcasts and laboratory broadcast simulations recorded from seven FM radio announcers (4 male, 3 female). In this corpus, a majority of paragraphs are annotated with the orthographic transcription, phone alignments, part-of-speech tags and prosodic labels.

Our first experiment investigated the importance of including large syntactic context and the phrase opening/closing information for syntactic-prosodic modeling. The training accuracies of the syntactic-prosodic model under different conditions, obtained from speaker F2B, are listed in Table 1, where n denotes the size of the syntactic information window (as we have discussed before), POC stands for the phrase opening/closing information returned by the Charniak parser. The results in Table 1 indicate that: 1. large syntactic context is important for both accent and boundary prediction; 2. POC is more useful in small context (n=1, 3) than in large context (n=5, 7).

To estimate the performance of all three types of models introduced in section 2 on this corpus, we applied a leave-one-speaker-out strategy. Data used in the experiments are extracted from 4

n	with POC		without POC	
	Acc. (%)	IPB (%)	Acc. (%)	IPB (%)
1	83.7	89.7	83.6	82.0
3	85.0	91.9	84.9	91.4
5	86.2	93.0	86.0	91.9
7	86.1	93.0	86.4	92.5

Table 1. The accent (Acc.) and the IPB prediction accuracy (%) of the ANN syntactic-prosodic model under various conditions: n=1,3,5,7, with or without phrase opening/closing information (POC), trained and tested on F2B.

Speakers	F2B	F1A	M1B	M2B
# Utterances	164	51	38	33
# Words	14844	3098	3366	2363
# Accents	6345	1382	1500	1061
# IPBs	2744	497	445	409

Table 2. The number of utterances, number of words, number of accents and number of intonational phrase boundaries (IPBs) for the 4 speakers used in our experiment.

speakers: F1A, F2B, M1B and M2B (where F/M designates female/male speakers). For each experiment, we used data from one speaker for test and the other three for training. F2B was never left-out because it contains the most data. The statistics of the speakers are listed in Table 2 and the average (weighted by number of words in each speaker) recognition results are listed in Table 3.

As shown in Table 3, the acoustic-prosodic model alone (AP) results are slightly worse than Wightman's results (84% for accent and 71% for IPB). However, our task is more difficult since our training set contains no utterance spoken by the test speaker; whereas Wightman's training and test set were formed by randomly dividing 2/3 of the data for training and the remaining 1/3 for testing with no speaker distinction. On the other hand, since our GMM-based acoustic model is simpler than Wightman's decision tree acoustic model both in the structure and in the dimensionality of input features (all GMMs in our experiments consist of 3 MGs), slightly worse results are expected. An advantage of our acoustic model is that it may provide better generalizability to unseen data (especially data from unseen speakers) than decision tree models because it is less likely to be overtrained due to its structural simplicity.

The syntactic-prosodic model alone (SP) results are very good. Especially, the IPB recognition accuracy has reached 90% which is 7% better than the chance level 83%. Accent can also be predicted from syntax with an 82.7% accuracy. The coupled model achieved accent recognition accuracy of 84.2% and IPB recognition accuracy of 93%, approaching the agreement rate between different human labelers (85-95% for accent, 95-98% for IPB using ToBI) for both accent and IPB recognition.

	Accent	IPB
SP	82.67	90.09
AP	77.34	68.15
ASP	84.21	93.07

Table 3. The averaged accent and IPB recognition accuracy (%) for the syntactic-prosodic model alone (SP), acoustic-prosodic model alone (AP) and the coupled model (ASP) on the leave-one-speaker-out task on the Radio News Corpus.

4. CONCLUSIONS

In this paper, we developed an ANN-based syntactic-prosodic model, which can be used to assign prosody labels from text for speech synthesis, and a GMM-based acoustic-prosodic model that can be coupled with the syntactic-prosodic model to improve prosody recognition accuracy for speech understanding systems. Our experiments on Radio News Corpus show that ANN is very effective in learning the stochastic mapping from the syntactic representation of word strings to prosody labels, with an accuracy of 82.7% for pitch accent labeling and 90.1% for intonational phrase boundary (IPB) labeling. When speech and reasonably accurate phoneme transcriptions are given, a GMM-based acoustic-prosodic model, coupled with the syntactic-prosodic model as a maximum likelihood recognizer, can achieve the pitch accent recognition accuracy of 84% and IPB recognition accuracy of 93% in a leave-one-speaker-out task. These results have considerably exceeded previous reported results on the same corpus and are approaching the agreement rate among human labelers.

5. REFERENCES

[1] J. Hirschberg, "Pitch Accent in Context: Predicting Intonational Prominence from Text," *Artificial Intelligence*, vol. 63 no. 1-2, 1993.

[2] S. Arnfield, "Prosody and syntax in corpus based analysis of spoken English," Ph.D. dissertation, University of Leeds, Dec. 1994.

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

[4] M. E. Beckman and G. A. Elam, "Guidelines for ToBI labelling," 1994, http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/singer_tobi.html.

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

[6] R. Kompe, "Prosody in speech understanding systems," *Lect. Notes in Artificial Intelligence*, Springer-Verleg, 1307:1-357, 1997.

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

[8] E. Charniak, "A maximum-entropy-inspired parser," in *Proceedings of NAACL*, 2000

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

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