Automatic Prosody Prediction and Detection with Conditional Random Field (CRF) Models

Yao Qian, Zhizheng Wu*, Xuezhe Ma*, Frank Soong
Microsoft Research Asia, Beijing, China
(yaoqian, frankkps)@microsoft.com

Abstract— **While the current TTS systems can deliver quite acceptable segmental quality of synthesized speech for voice user interface applications, its prosody is still perceived by users as "robotic" or not expressive. In this paper, we investigate how to improve TTS prosody prediction and detection. Conditional Random Field (CRF), a discriminative probabilistic model for the labeling the sequential data, is adopted. Rich syntactic and acoustic, contextual features are used in building the CRF models. Experiments performed on Boston University Radio Speech Corpus show that CRF models trained on our proposed rich contextual features can improve the accuracy of prosody prediction and detection in both speaker-dependent and speaker-independent cases. The performance is either comparable or better than the best reported results.**

Keywords-Prosody; Prosody label prediction; Prosody event detection; CRF; acoutic; syntatic

I. Introduction

Prosody refers to the rhythm, stress, and intonation of speech. The acoustic cues of prosody include variations in syllable duration, loudness and pitch and they play an important perceptual role in human speech communication. For tonal language like Mandarin, the variation of pitch has an additional lexical meaning, which is used to distinguish one lexical entity from the other. For non-tonal languages like English, pitch accent can express the focus and give additional information. Boundary tone can differentiate a simple statement sentence from a yes-no question. Prosody also carries the information of a speaker's emotion state, e.g., exciting or bored. However, it is still fairly enigmatic in linguistic science research about how a person decides which words to accentuate; what factors constrain the accent placement and what function an accent serves in conveying message. For speech synthesis, it is a great challenge to predict the correct prosody labels from unrestricted text or detect prosody event jointly with acoustic features.

Lots of studies on prosody prediction and detection have been done to investigate the syntactic, semantic, and discourse/pragmatic structure and its relevance to prosody generation in speech production. The syntactic cues like part-of-speech, syllable identity, syllable stress and their contextual counterparts are commonly used for prosody label prediction [1-10]. Pitch, segmental duration, intensity and other acoustic correlates are generally adopted as assistants to prosody events detection [1-10]. Some informative features such as word class, word predictability (language model) and term frequency, inverse document frequency (TF-IDF) have been successfully applied to prosody prediction [5, 6]. To model prosody, many machine learning techniques have also been investigated. Decision tree, neural network, Gaussian mixture model, and hidden Markov models are generally applied to model prosody. In [7], different classifiers were studied for prosody event detection. Among them, the support vector machine (SVM) has been claimed to perform better than other classifiers to model syntactic evidence. In [8], it got better results over the classification and regression trees (CART) by using bagging and boosting ensemble learning approaches. Recently, discriminate training approaches like maximum entropy model and conditional random field (CRF) are tried for prosody prediction and detection and good performance are reported [9, 10].

In this paper, our approach to prosody prediction is based upon CRF modeling with rich, phonetic, text based or syntactic, and acoustic features. The rest of paper is organized as follows. In Section II, the corpus and tasks for our experiments are reviewed. The principle of CRF and the feature sets used for our CRF modeling are introduced in Section III and IV, respectively. Experimental evaluations and their results are presented in Section V. Conclusions are given in Section VI.

II. Corpus Description and Tasks

We use speech data recorded by professional radio announcers from Boston University Radio Speech Corpus (BURSC) [11] for this study. The corpus is a database annotated with the tones and break indices (ToBI) [12] prosodic annotation system. Subsets of the corpus are automatically labeled with phonetic alignments, part-of-speech tags and hand-labeled prosodic markers. For our experiments, we use a speaker-independent speech recognizer to obtain the word, syllable and phone boundaries by force alignment. The statistics of data for the six speakers is listed in Table 1.

In a TOBI framework, there are eight pitch accent types, two intermediate phrase boundary tones, four intonational phrase boundary tones and five break indices, where 3 for intermediate phrase boundary and 4 for intonation phrase boundary. We only focus on pitch accent and break prediction /detection in our experiments since the phrase boundary tone is

*Work performed as an intern in the Speech Group, Microsoft Research Asia

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

less variable in broadcast news style corpus like BURSC. In order to streamline the task and to make the output result more useful, we follow earlier approaches [8,9,12] by collapsing the TOBI labels as follows:

- Pitch Accent

L*	L*+!H	L*+H	->	L*
!H*	H+!H*	L+!H*	->	!H*
H*	L+H*		->	H*
Unaccent			->	X

- Break

break ending with a punctuation (, : . ; ?)	->	#
break (3,4) within a sub-sentence	->	1
Otherwise	->	0

We assume it is both natural and acceptable to assign a break at a punctuation mark, which is used to divide a long sentence into several sub-sentences (phrases). It is then critical to predict breaks reliably within a sub-sentence for real applications. Therefore, we only predict and detect breaks within a sentence or sub-sentence where no punctuation marks are given in our experiments.

Table 1. The statistics of data used in our experiments

	Female			Male		
	F1A	F2B	F3A	M1B	M2B	M3B
#Utterances	76	147	33	71	51	23
#Sentences	248	582	154	258	213	104
# Words	4386	11110	2732	5015	3608	1936
# Syllables	7084	18027	4348	7977	5867	3085
# Breaks	1217	3422	754	1231	992	446
# Accent	2513	6159	1003	2784	2017	1091

There are two tasks performed in this study: predicting prosody labels from raw text and detecting prosody events from raw text along with corresponding acoustic signals. Prediction task is the same as to generate prosodic labels in a text-to-speech (TTS) system, while detection task is employed to automatically annotate the training data of TTS since the prosody annotation is a time-consuming work and human annotations tend to be inconsistent annotations even among experienced annotators. Each task is carried out on two data sets: F2B, a speaker-dependent (SD) set, and speaker-independent (SI) set.

III. Prosody Modeling with CRF

The prosodic prediction and detection can be formulated as a problem of sequential labeling. Given an utterance (a sequence of words or syllables) with N tokens, the task is to find the most probable sequence of tone and break indices for N junctures after every token

$$\hat{\vec{J}} = \arg\max_{\vec{J}} \{P(\vec{J} \mid \vec{T})\} \tag{1}$$

For example, a possible break assignment of utterance “It functions like an electronic officer.” is after the word “functions”, the corresponding input/output vectors are:

$$\vec{T} = <"It","functions","like","an","electronic","officer",".">$$

$$\vec{J} = <"0","1","0","0","0","\#">$$

Here punctuations are not considered as “word”, while the word precedes a punctuation is assigned a break “#”. 0 in the break assignment vector represents a non-break juncture and 1 represents a break juncture. Similar input/output vectors are defined in pitch accent problem.

The state-of-art model for perform sequential labeling is conditional random field (CRF) [13], a graphical model for computing the conditional probability $p(\vec{y} \mid \vec{x})$ of a label sequence $\vec{y}$, $\vec{y} = y_1, y_2, \ldots, y_n$, given observation sequence $\vec{x}$, $\vec{x} = x_1, x_2, \ldots, x_n$. Linear chain CRF, a special form of CRF, is used in this study. It has the form:

$$p_{\vec{\lambda}}(\vec{y} \mid \vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \exp(\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_j f_j(y_{i-1}, y_i, \vec{x}, i)) \tag{2}$$

where i is the position of input label sequence; Each feature function $f_j(y_{i-1}, y_i, \vec{x}, i)$ is either a state function $s(y_i, \vec{x}, i)$ or a transition function $t(y_{i-1}, y_i, \vec{x}, i)$; λ_j are parameters to be estimated from training data and $Z_{\vec{\lambda}}(\vec{x})$ is for normalization. CRFs are usually trained by maximizing the log-likelihood over a given training set. Limited memory BFGS [14] is adopted in training to solve this unconstrained convex optimization problem.

Prosody modeling can be seen as relational learning [15]. There are two characteristics in relational data. One is that the labels in a label sequence ($\vec{y}$), which we want to model, have statistical dependencies. The other one is that each label often has a rich set of features ($\vec{x}$) that are useful for classification. The main advantage of CRF is that it can include rich, overlapping features. It incorporates the dependency among observations and aims to solve the long distance dependency problem. In CRF, features based on rich relationships between input and output vectors are easily incorporated.

IV. Features for Prosody Modeling

We investigate features for prosody modeling. Features for both pitch accent and break modeling include features extracted from raw text and acoustic features extracted from speech signals. There are total four features sets investigated:

- Text feature set for pitch accent:

  - The phonetic form of the current syllable, the previous two syllables, and the following two syllables;
  - Binary indicators to label whether each of the current, previous, and following syllables are lexically stressed;
  - The position of the current, previous and next syllable in a word;
  - Part-of-speech of the current, two previous words and the following two words;
  - The current word, the two previous word and the following two words;
  - A composite features made up of part-of-speech and stress of the current syllable;
  - A composite features made up of part-of-speech and the current syllable vowel;
  - A composite features made up of current word and stress of the current syllable

- Acoustic feature set for pitch accent:
  - F0 mean of previous two syllables, current and next two syllables;
  - Delta and delta-delta F0 mean of the current syllable;
  - F0 Slope of previous two syllables, current and next two syllables;
  - Delta and delta-delta F0 slope of the current syllable;
  - A composite features made up of F0 mean and F0 slope;
  - Range of the current syllable's F0

- Text feature set for break:
  - Word
  - Number of syllables in the current word
  - POS
  - Punctuation
  - Function word
  - Capital
  - Bigram LM
  - Phrase dictionary
  - Above features' quin-context and combination

- Acoustic feature set for break:
  - Silence after word,
  - Duration of last syllable
  - Duration of last stressed syllable
  - Above features' quin-context and combination

We propose some new features for prosody modeling listed in the above feature sets. Prosody is a supra-segmental feature. It is laid on top of groups of segments. The relative increase or decrease of F0 in comparing with neighbouring syllables should be a relevant indicator of pitch accent. We add the dynamic features (delta and delta-delta) of F0 mean and slope of a syllable to pitch accent modeling. In addition, according to our analysis for the data annotated by break, we find that current word and next word with a high bigram language model (LM) probability, allows no break in between (i.e., read as a continues chunk without break), as shown in Figure 1. However, this phenomenon is much more distinctive in content word pairs, not for function word pairs. In addition, a heuristic assumption is that words within a phrase usually can't assign a break. It is also confirmed by native speakers' speaking habit. Therefore, we propose to use content word pair bigram probabilities and a phrase dictionary in break modeling.

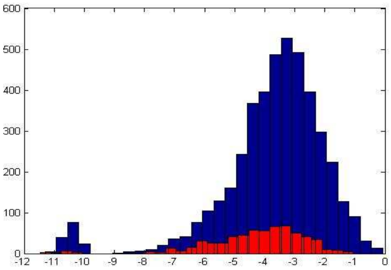

Fig 1. The histogram of break (smaller one) and non-break (larger one) along with two content words bigram probabilities (horizontal axis).

Acoustic features like F0 are highly variable. Speaker difference in F0 can be affected by many factors, e.g. age, gender and personal style. Even for the same speaker, the range of F0 can change from utterance to utterance. Normalization is then necessary to minimize undesirable fluctuations. The F0 value of each frame is normalized (divided) by the mean of F0 of the whole sentence and the duration is normalized (divided) by the mean syllable duration averaged over all speech data of the speaker.

V. Expereimtns and Results

We use F2B's data for speaker-dependent (SD) task. The data proportion for training, developing and testing is roughly 4:1:1. The experiments for speaker-independent (SI) are carried out with a cross-validation procedure, i.e., training on multiple speakers, testing on a held-out speaker, and reporting the average across the whole tests. Since F2B has much more data than other speakers, she is always used in the training set. We also leave 20% data out from training set for development.

CRF tool we used is CRF++ toolkit [16], a simple, customizable, and open source implementation of CRFs for segmenting/labeling sequential data. L-BFGS, a quasi-Newton algorithm for large scale numerical optimization problem, is used for training and L2 norm is adopted for regularization. The features listed in Section IV all contribute to prosody prediction and detection, confirmed by our brute force testing with a super feature set. The CRF training control parameters, e.g., the cut-off threshold for the features, are optimized with the development set. CRF tools in natural language processing area don't support continuous features. The acoustic features are then quantized into brackets which are also optimized with the development set.

To compare with results in other references [8,9], we use accuracy as a performance evaluation measure and four speakers' (F1A, F2B,M1B and M2B) cross-validation procedure for SI model. The results of pitch accent prediction and detection are shown in Table 2, where Sun's [8] and Levow's [9] results were the best reported results with the same data set and pitch accent types, as far as we know. The performance of our pitch accent detection outperforms the Sun's SD model and Levow's SI model by 0.78% and 3.28%, respectively. By analyzing the results further, we find most improvements are from our new acoustic features: dynamic features of F0 mean and slope of syllables.

Table 2. The results for pitch accent prediction and detection.

Accuracy (%) Features	SD		SI	
	Sun's [8]	Our	Levow's [9]	Our
Text	84.71	81.46	76.21	77.49
Acoustic	80.60	83.50	77.06	77.38
Text+Acoustic	87.17	87.95	79.65	82.93

In evaluating the break model, we use two metrics: 1) precision, recall and f-score for break prediction within a sub-sentence or sentence without punctuations; 2) whole accuracy for break and non-break prediction. The results of break prediction for SI and SD tasks are shown in Tables 3 and 4, where our new features: phrase dictionary and content word LM probability can improve the F-score from 70.2% to 72.0%. We also show the evaluation results by using different ground truth in Table 4. It is observed from the database that the break assignment can be highly idiosyncratic and even random. As a result, we count the predicted break as correct it matches any one of four speakers. The four speakers uttered the same sentence but with different breaks. A typical example:

Predictor: It functions | like an electronic probation officer.
F2B: It functions | like an electronic | probation officer.
F1A: It functions | like an electronic | probation officer.
M1B: It functions | like an electronic probation officer .
M2B: It functions like an electronic probation officer .

The same sentence can be assigned by non, one or two breaks by different speakers. From this point of view, we find the performance of our break model is pretty good and it can be applied to a real TTS system.

Table 3. The results of break prediction for SI task.

Text Features	Precision (%)	Recall(%)	F-score(%)
Baseline	69.0	71.8	70.2
Our features	71.2	72.9	72.0

Table 4. The results of break prediction for SD task.

Ground Truth	Precision (%)	Recall(%)	F-score(%)
F2B	75.6	76.6	76.1
Four speakers	85.6	86.7	86.1

To compare with the results in reference [7], the newest report on prosody modeling with BURSC data, we also use the accuracy of break and non-break on the cross-validation procedure of six speakers to evaluate the performance of break model. The results are shown in Table 5 and they are comparable or better than those in reference [7].

Table 5: The Accuracy of break prediction and detection for SI task.

Accuracy (%) Features	5 fold (six speakers)	
	Jeon's [7]	Our
Text	89.76	91.18
Acoustic	84.89	84.78
Text+Acoustic	91.06	92.11

VI. Conclusions

Prosody modeling is a critical component for generating expressive TTS speech. We use CRF for prosody modeling with rich syntactic and acoustic, contextual features. The experimental results on BURSC corpus show that our prosody model can achieve 82.93% and 92.11% accuracy for pitch accent and break detection, respectively.

References

[1] Y. Qian, M Chu, H. Peng. "Segmenting unrestricted Chinese text into prosodic words instead of lexical words", In Proc. of ICASSP, 2001.

[2] Y. Qian and F. Chen, "Assigning phrase accent to Chinese text-to-speech system", In Proc.of ICASSP, 2002 .

[3] X Sun, TH Applebaum. "Intonational phrase break prediction using decision tree and n-gram model." In Proc. of Eurospeech, 2001.

[4] P. Koehn, S. Abney, J Hirschberg, M Collins. "Improving intonational phrasing with syntactic information." In Proc. of ICASSP, 2000.

[5] J. hirschberg. Pitch Accent in Context: Predicting Intonational Prominence from Text. Artificial Intelligence 63 (1-2),1993.

[6] S. pan and K. R. McKeown. "Word Informativeness and Automatic Pitch Accent Modeling", In Proc. Of EMNLP/VLC,1999.

[7] J. H. Jeon and Y. Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features", In proc. of ICASSP, 2009.

[8] X. Sun, "Pitch accent prediction using ensemble machine learning", In Proc. of ICSLP, 2002.

[9] G.-A. Levow. "Automatic prosodic labeling with conditional random fields and rich acoustic features". In Proc. of IJCNLP, 2008.

[10] V. Rangarajan, S. Narayanan, S. Bangalore. "Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework", IEEE Trans on ASLP, 2008.

[11] M. Ostendorf, PJ Price, S Shattuck-Hufnagel. The Boston University radio news corpus.

[12] M. Ostendorf and K. Ross, A multi-level model for recognition of intonation labels, In Y. Sagisaka, N. Campbell and N. Higuchi, editors, *Computing Prosody*, pp. 291-308, 1997.

[13] J. Lafferty, A. McCallum and F. Pereira, "Conditional ramdon fields: Probabilistic modles for segmenting and labeling sequence data", *In Proc. of 18 th ICML*, pp. 282-289, 2001.

[14] J. Necedal, Undating quasi-newton matrices with limited storage, *Mathematics of Computation*, Vol. 35, pp. 773-782, 1980.

[15] C. Sutton and A. McCallum, An Introduction to Conditional Random Fields for Relational Learning, *Book chapter in Introduction to Statistical Relational Learning*. Edited by Lise Getoor and Ben Taskar. MIT Press. 2006.

[16] CRF++: Yet Another CRF toolkit , http://crfpp.sourceforge.net/.