Automatic Prosodic Labeling with Conditional Random Fields and Rich Acoustic Features

Gina-Anne Levow
University of Chicago
Department of Computer Science
1100 E. 58th St.
Chicago, IL 60637 USA
levow@cs.uchicago.edu

Abstract

Many acoustic approaches to prosodic labeling in English have employed only local classifiers, although text-based classification has employed some sequential models. In this paper we employ linear chain and factorial conditional random fields (CRFs) in conjunction with rich, contextually-based prosodic features, to exploit sequential dependencies and to facilitate integration with lexical features. Integration of lexical and prosodic features improves pitch accent prediction over either feature set alone, and for lower accuracy feature sets, factorial CRF models can improve over linear chain based prediction of pitch accent.

1 Introduction

Prosody plays a crucial role in language understanding. In addition to the well-known effects in tone languages such as Chinese, prosody in English also plays a significant role, where pitch accents can indicate given/new information status, and boundary tones can distinguish statements from yes-no questions. However, recognition of such prosodic features poses significant challenges due to differences in surface realization from the underlying form. In particular, context plays a significant role in prosodic realization. Contextual effects due articulatory constraints such maximum speed of pitch change (Xu and Sun, 2002) from neighboring syllables and accents can yield co-articulatory effects at the intonational level, analogous to those at the segmental level. Recent phonetic research (Xu, 1999; Sun, 2002; Shen, 1990) has demonstrated the importance of coarticulation for tone and pitch accent recognition. In addition context affects interpretation of prosodic events; an accent is viewed as high or low relative to the speaker's pitch range and also relative to adjacent speech.

Some recent acoustically focused approaches (Sun, 2002; Levow, 2005) to tone and pitch accent recognition have begun to model and exploit these contextual effects on production. Following the Parallel Encoding and Target Approximation (PENTA) (Xu, 2004), this work assumes that the prosodic target is exponentially approached during the course of syllable production, and thus the target is best approximated in the later portion of the syllable. Other contextual evidence such as relative pitch height or band energy between syllables has also been employed (Levow, 2005; Rosenberg and Hirschberg, 2006). Interestingly, although earlier techniques (Ross and Ostendorf, 1994; Dusterhoff et al., 1999) employed Hidden Markov Models, they did not explicitly model these coarticulatory effects, and recent approaches have primarily employed local classifiers, such as decision trees (Sun, 2002; Rosenberg and Hirschberg, 2006) or Support Vector Machines (Levow, 2005).

Another body of work on pitch accent recognition has focused on exploitation of lexical and syntactic information to predict ToBI labels, for example for speech synthesis. These approaches explored a range of machine learning techniques from local classifiers such as decision trees (Sun, 2002) and RIPPER (Pan and McKeown, 1998) to sequence models such as Conditional Random Fields

(CRFs)(Gregory and Altun, 2004) more recently. The systems often included features that captured local or longer range context, such as n-gram probabilities, neighboring words, or even indicators of prior mention. (Chen et al., 2004; Rangarajan Sridhar et al., 2007) explored the integration of based prosodic and lexico-syntactic evidence in GMM-based and maximum entropy models respectively.

Here we explore the use of contextual acoustic and lexical models within a sequence learning framework. We analyze the interaction of different feature types on prediction of prosodic labels using linear-chain CRFs. We demonstrate improved recognition by integration of textual and acoustic cues, well-supported by the sequence model. Finally we consider the joint prediction of multiple prosodic label types, finding improvement for joint modeling in the case of feature sets with lower initial performance.

We begin by describing the ToBI annotation task and our experimental data. We then discuss the choice of conditional random fields and the use of linear chain and factorial models. Section 4 describes the contextual acoustic model and text-based features. Section 5 presents the experimental structure and results. We conclude with a brief discussion of future work.

2 Data

We employ a subset of the Boston Radio News Corpus (Ostendorf et al., 1995), employing data from speakers f1a, f2b, m1b, and m2b, for experimental consistency with (Chen et al., 2004; Rangarajan Sridhar et al., 2007). The corpus includes pitch accent, phrase and boundary tone annotation in the ToBI framework (Silverman et al., 1992) aligned with manual transcription and manual and automatic syllabification of the materials. Each word was also manually part-of-speech tagged. The data comprises over forty thousand syllables, with speaker f2b accounting for just over half the data. Following earlier research (Ostendorf and Ross, 1997; Sun, 2002), we collapse the ToBI pitch accent labels to four classes: unaccented, high, low, and downstepped high for experimentation, removing distinctions related to bitonal accents. We also consider the binary case of distinguishing accented from unaccented syllables, (Gregory and Altun, 2004; Rosenberg and Hirschberg, 2006; Ananthakrishnan and Narayanan, 2006). For phrase accents and boundary tones, we consider only the binary distinction between phrase accent/no phrase accent and boundary tone/no boundary tone.

All experiments evaluate automatic prosodic labeling at the syllable level.

3 Modeling with Linear-Chain and Factorial CRFs

Most prior acoustically based approaches to prosodic labeling have used local classifiers. However, on phonological grounds, we expect that certain label sequences will be much more probable than others. For example, sequences of multiple high accents are relatively uncommon in contrast to the case of an unaccented syllable preceding an accented one. This characteristic argues for a model which encodes and exploits inter-label dependencies. Furthermore, under the ToBI labeling guidelines, the presence of a boundary tone dictates the co-occurrence of a phrase accent label. To capture these relations between labels of different types, we also consider factorial models.

Conditional Random Fields (Lafferty et al., 2001) are a class of graphical models which are undirected and conditionally trained. While they can represent long term dependencies, most applications have employed first-order linear chains for language and speech processing tasks including POS tagging, sentence boundary detection (Liu et al., 2005), and even text-oriented pitch accent prediction(Gregory and Altun, 2004). The models capture sequential label-label relations, but unlike HMMs, the conditionally trained model can more tractably support larger text-based feature sets. Factorial CRFs (Sutton, 2006; McCallum et al., 2003) augment the linear sequence model with additional cotemporal labels, so that multiple (factors) labels are predicted at each time step and dependencies between them can be modeled. Examples of linear-chain and factorial CRFs appear in Figure 1. In the linear chain example, the f_i items correspond to the features and the y_i to labels to be predicted, for example prosodic and text features and pitch accent labels respectively. The vertical lines correspond to the dependencies

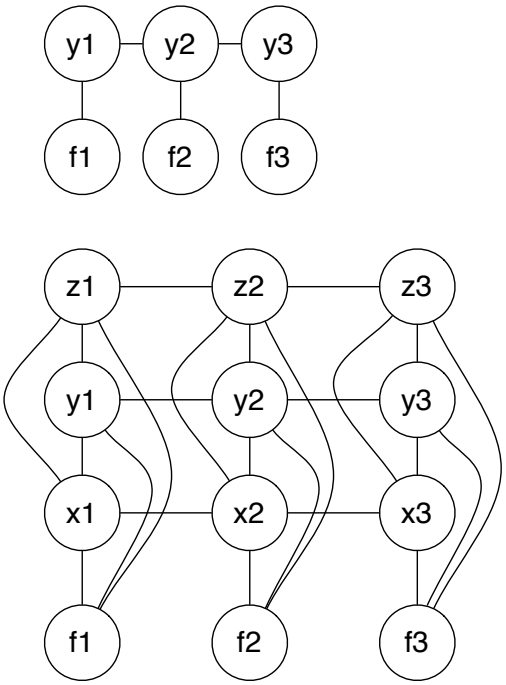

Figure 1: Linear-chain CRF (top) and Two-level Factorial CRF (bottom).

between the features and labels; the horizontal lines indicate the dependencies between the labels in sequence. In the factorial CRF example, the f_i again represent the features, while the x_i, y_i, and z_i represent the boundary tone, phrase accent, and pitch accent labels that are being predicted. The horizontal arcs again model the sequential bigram label-label dependencies between labels of the same class; the vertical arcs model the dependencies between both the features and labels, and bigram dependencies between the labels of each of the different pairs of factors. Thus, we jointly predict pitch accent, phrase accent, and boundary tone and, the prediction of each label depends on the features, the other labels predicted for the same syllable, and the sequential label of the same class. So, pitch accent prediction depends on the features, pitch accent predicted for the neighboring syllable, and phrase and boundary tone predictions for the current syllable.

We employ the Graphical Models for Mallet (GRMM) implementation (Sutton, 2006), adapted to also support the real-valued acoustic features required for these experiments; in some additional contrastive experiments on zero order models, we also employ the Mallet implementation (McCallum, 2002). We employ both linear chain and three-level factorial CRFs, as above, to perform prosodic labeling.

4 Feature Representation

We exploit both lexical and prosodic features for prosodic labeling of broadcast news speech. In particular, in contrast to (Gregory and Altun, 2004), we employ a rich acoustic feature set, designed to capture and compensate for coarticulatory influences on accent realization, in addition to word-based features.

4.1 Prosodic Features

Using Praat's (Boersma, 2001) "To pitch" and "To intensity" functions and the phoneme, syllable, and word alignments provided in the corpus, we extract acoustic features for the region of interest. This region corresponds to the syllable nucleus in English. For all pitch and intensity features, we compute per-speaker z-score normalized log-scaled values.

Recent phonetic research (Xu, 1997; Shih and Kochanski, 2000) has identified significant effects of carryover coarticulation from preceding adjacent syllable tones. To minimize these effects consistent with the pitch target approximation model (Xu et al., 1999), we compute slope features based on the second half of this region, where this model predicts that the underlying pitch height and slope targets of the syllable will be most accurately approached.

For each syllable, we compute the following local features:

- pitch values at five points evenly spaced across the syllable nucleus,
- mean and maximum pitch values,
- slope based on a linear fit to the pitch contour in the second half of the region, and
- mean and maximum intensity.

We consider two types of contextualized features as well, to model and compensate for coarticulatory effects from neighboring syllables. The first set of features, referred to as "extended features", includes the maximum and mean pitch from adjacent

syllables as well as the nearest pitch points from the adjacent syllables. These features extend the modeled tone beyond the strict bounds of the syllable segmentation. A second set of contextual features, termed "difference features", captures the change in feature values between the current and adjacent syllables. The resulting feature set includes:

- mean, maximum, and last two pitch values from preceding syllable,
- mean, maximum, and first value from following syllable, and
- differences in pitch mean, pitch maximum, pitch of midpoint, pitch slope, intensity mean, and intensity maximum between the current syllable and the preceding syllable, and between the current syllable and the following syllable.

Finally, we also employ some positional and durational features. Many prosodic phenomena are affected by phrase or sentence position; for example, both pitch and intensity tend to decrease across an utterance, and pitch accent realization may also be affected by cooccurring phrase accents or boundary tones. As syllable duration typically increases under both accenting and phrase-final lengthening, this information can be useful in prosodic labeling. Finally, pause information is also associated with prosodic phrasing. Thus, we include following features:

- two binary features indicating initial and final in a pseudo-phrase, defined as a silence-delimited interval,
- duration of syllable nucleus, and
- durations of pause preceding and following the syllable.

In prior experiments using support vector machines (Levow, 2005), variants of this representation achieved competitive recognition levels for both tone and pitch accent recognition.

4.2 Text-based Features

We employ text-based models similar to those employed by (Sun, 2002; Rangarajan Sridhar et al., 2007). For each syllable, we capture the following manually annotated features:

- The phonetic form of the current syllable, the previous two syllables, and the following two syllables,
- binary values indicating whether each of the current, previous, and following syllables are lexically stressed,
- integer values indicating position in a word of the current, previous, and following syllables,
- the current word, the two previous words, and the two following words, and
- the POS of the current word, of the two previous words, and of the two following words.

These features capture information about the current syllable and its lexico-syntactic context, that have been employed effectively in prosodic labeling of pitch accent, phrase accent, and boundary tone.

5 Experiments

We explore a range of issues in the experiments reported below. We hope to assess the impact of feature set and acoustic and text-based feature integration in the Conditional Random Field models. We compare their individual effectiveness as well as the effect of combined feature sets on labeling. In particular, we consider both the binary accented/unaccented assignment task for pitch accent and the four way - high/downstepped high/low/unaccented - contrast to compare effectiveness in problems of different difficulty. We further consider the effect of sequence and factorial modeling on pitch accent recognition. All experiments are conducted using a leave-one-out evaluation procedure following (Chen et al., 2004), training on all but one speaker and then testing on that held-out speaker, reporting the average across the tests on held-out data. Because speaker f2b contributes such a large portion of the data, that speaker is never left out.

On this split, the best word-based accuracy incorporating both prosodic and lexico-syntactic information in a maximum entropy framework is 86.0% for binary pitch accent prediction and 93.1% for

recognition of boundary status (Rangarajan Sridhar et al., 2007). For syllable-level recognition on this dataset, results for speaker-independent models reach slightly over 80% for binary pitch accent detection and 88% for boundary detection. Speaker dependent models have achieved very high accuracy; over 87% on speaker f2b was reported by (Sun, 2002) for the four-class task.

5.1 Explicit Prosodic Context Features and Sequence Models

We first assess the role of contextual prosodic features for pitch accent recognition and their interaction with sequence models. To minimize interaction effects, we concentrate on recognition with prosodic features alone on the challenging four-way pitch accent problem. As described above, we augmented the local syllable-based prosodic features with contextual features associated with the preceding and following syllables. We ask whether the use of contextual features improves recognition, and, if so, which type of context, preceding or following, has the greatest impact. We also ask whether the CRF models provide further improvements or can partially or fully compensate for the lack of explicit context features. To evaluate this impact, we compute four-way pitch accent recognition accuracy with no context features, after adding preceding context, after adding following context, and with both. We also contrast zero order and first order linear chain CRFs for these conditions. We find that modeling preceding context yields the greatest improvement. This finding is consistent with findings in recent phonetic research that argue for a larger role of carryover coarticulation from preceding syllables than of anticipatory coarticulation with following syllables. Furthermore, sequence modeling in the CRF also improves results, across the explicit context feature conditions, with improvements being most pronounced in cases with less effective explicit prosodic contextual features. Results for prosodic features alone appear in Table 1. In a side experiment with these prosodic features, we also briefly explored higher-order models, but no improvement was observed.

We also assess the impact of this richer contextualized prosodic feature set both alone and in conjunction with the full text-based feature set, in the full factorial CRF framework. We compare results for pitch accent identification in both the two-way and four-way conditions with no context and with the full ensemble of prosodic features. We find no difference for the two-way, all features condition for which text-based features perform well alone. However, for the prosody only cases and the more challenging four-way task with all features, contextual information yields improvements, demonstrating the utility of this richer, contextualized prosodic feature representation. These contrasts appear in Table 2.

		No Context	Full Context
Prosody	Two-way	78.9%	80.8%
Only	Four-way	74.2%	78.2%
All	Two-way	86.2%	86.2%
Features	Four-way	79%	79.7%

Table 2: Impact of context prosodic features with prosody alone and all features

5.2 Prosodic and Text-based Features

We continue by contrasting effectiveness of different feature sets in the basic linear-chain CRF case for pitch accent recognition. Table 3 presents the results for prosodic, word-based, and combined features sets in both the two-way and four-way classification conditions. Overall accuracy is quite good; in all cases, results are well above the 65% most common class assignment level, and the best results (86.2%) outperform any previously published speaker independent syllable-based results on this dataset. Overall results and contrasts are found in Table 3.

It is clear that the two feature sets combine very effectively. In the 4-way pitch accent task, the combined model yields a significant 1.5% to 2.5% increase over the strong acoustic-only model. In contrast, in the binary task, both the overall effectiveness of the text-based model and its utility in combination with the acoustic features are enhanced, yielding a much higher individual and combined accuracy rate. This contrast can be explained by the fact that the word features, such as part of speech, identify items that, as a class, are likely to be accented rather then being strongly associated with a particular tone category. The type of accent is likely

	No Context	Preceding	Following	Both
Zero order	70.5%	75.2%	71.8%	76.4%
First order	74.2%	75.5%	73.7%	77.1%

Table 1: Prosodic Context Features and CRFs

		Acoustic	Text	Text&Acoustic
Linear-Chain	Two-way	79.48%	84.88%	86.1%
	Four-way	77.06%	76.21%	79.65%
Factorial CRF	Two-way	80.76%	84.74%	86.2%
	Four-way	78.22%	77.46%	79.71%

Table 3: Pitch Accent Classification with Linear-Chain (top) and factorial CRFs (bottom) , using Acoustic-only, Text-based-only, and Combined Features. Results for two- and four-way pitch accent prediction are shown.

best determined by acoustic contrast, since accent type is closely linked to pitch height, and the local context and acoustic features serve to identify which accentable words are truly accented. Thus, in the binary task, the text-based features combine most effectively with the evidence from the acoustic features.

To contrast local classifiers with the linear chain model with text-based features, we trained a zero order classifier for the pitch accent prediction case and contrasted it with a comparable first-order linear-chain CRFs. Here for the binary accent recognition case, using only text-based information, we reach an accuracy of 84.3% for the history-free model, contrasted with an 85.4% level obtained with a comparable first-order model.[1]

5.3 Factorial CRF Framework

Finally we consider the effect of joint classification using the factorial CRF framework. Here, beyond just pitch accent assignment, we perform simultaneous assignment of pitch accent, phrase accent and boundary tone, where each label type corresponds to a factor, implementing the desired dependencies.[2]

[1]This comparison was computed using the original Mallet CRF package rather than GRMM, due to simpler zero order model support. This results in a small difference in the resulting scores.

[2]The features have not been tuned specifically for phrase account and boundary prediction, as explicit punctuation or sentence boundary features would have been useful but obvious giveaways. However, our goal is to assess the potential impact of combined classification, without excessive tuning.

The contrasts with the linear-chain model in terms of pitch accent prediction accuracy appear in Table 3. For the binary pitch accent condition, results are somewhat mixed. While there is a small but not significant decrease in accuracy for the text-only binary classification condition, the combined case shows little change and the prosodic case increases modestly. We note in one case that joint accuracy has risen when the pitch accent accuracy has dropped; we speculate that some additional compensation is needed to manage the effects of the severe class imbalance between the dominant "no-label" classes for phrase accent and boundary tone and other labels. For the four-way contrast between pitch accent types, we see small to modest gains across all feature sets, with the prosodic case improving significantly ($p < 0.025$). The best results for all but the two-way text-based classification task are found with the factorial CRF model.

For phrase accent and boundary tone prediction, phrase accent accuracy reaches 91.14%, and boundary tone accuracy 93.72% for all features. Text-based evidence is more effective than prosodic evidence in these cases, with text-based features reaching 91.06% for phrase accent and 92.51% and acoustic features only 86.73% and 92.37% respectively. However, little change is observed with the factorial CRF relative to a linear chain model trained on the same instances. The results for phrase accent and boundary tone recognition appear in Table 4.

	Phrase Accent	Boundary Tone
Prosodic	86.73%	92.37%
Text	91.06%	92.51%
Text+Prosodic	91.14%	93.72%

Table 4: Accuracy for phrase accent and boundary tone with prosodic, text-based, and combined features

6 Conclusion and Future Work

The application of linear-chain and factorial Conditional Random Fields for automatic pitch accent recognition and other prosodic labeling facilitates modeling of sequential dependencies as well as integration of rich acoustic features with text-based evidence. We plan to further investigate the modeling of dependencies between prosodic labels and the sequential modeling for acoustic features. Finally, we will also integrate prior work on subsyllable segmentation to identify the best approximation of the prosodic target with the CRF framework to produce a fine-grained sequence model of prosodic realization in context.

7 Acknowledgments

The author would like to thank Charles Sutton for providing the GRMM implementation, Andrew McCallum for the Mallet CRF implementation, and Siwei Wang and Sonja Waxmonsky for the modifications supporting real-valued features.

References

Sankaranarayanan Ananthakrishnan and Shrikanth Narayanan. 2006. Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling. In *Proceedings of ICSLP 2006*.

P. Boersma. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9–10):341–345.

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*.

K. Dusterhoff, A. Black, and P. Taylor. 1999. Using decision trees within the tilt intonation model to predict f0 contours. In *Proc. Of Eurospeech '99*.

Michelle Gregory and Yasemin Altun. 2004. Using conditional random fields to predict pitch accents in conversational speech. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 677–683, Barcelona, Spain, July.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML-2001)*.

Gina-Anne Levow. 2005. Context in multi-lingual tone and pitch accent prediction. In *Proc. of Interspeech 2005*.

Yang Liu, Andreas Stolcke, Elizabeth Shriberg, and Mary Harper. 2005. Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 451–458, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Andrew McCallum, Khashayar Rohanimanesh, and Charles Sutton. 2003. Dynamic conditional random fields for jointly labeling multiple sequences. In *NIPS*2003 Workshop on Syntax, Semantics, Statistics*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

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

M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. 1995. The Boston University radio news corpus. Technical Report ECS-95-001, Boston University.

Shimei Pan and Kathleen McKeown. 1998. Learning intonation rules for concept to speech generation. In *Proceedings of ACL/COLING-98*, pages 1003–1009.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. 2007. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 1–8, Rochester, New York, April. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2006. On the correlation between energy and pitch accent in read english speech. In *Proceedings of ICLSP 2006*.

K. Ross and M. Ostendorf. 1994. A dynamical system model for generating f0 for synthesis. In *Proceedings of the ESCA/IEEE Workshop on Speech Synthesis*, pages 131–134.

Xiao-Nan Shen. 1990. Tonal co-articulation in Mandarin. *Journal of Phonetics*, 18:281–295.

C. Shih and G. P. Kochanski. 2000. Chinese tone modeling with stem-ml. In *Proceedings of the International Conference on Spoken Language Processing, Volume 2*, pages 67–70.

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

Xuejing Sun. 2002. Pitch accent prediction using ensemble machine learning. In *Proceedings of ICSLP-2002*.

Charles Sutton. 2006. Grmm: A graphical models toolkit. http://mallet.cs.umass.edu.

Yi Xu and X. Sun. 2002. Maximum speed of pitch change and how it may relate to speech. *Journal of the Acoustical Society of America*, 111.

C.X. Xu, Y. Xu, and L.-S. Luo. 1999. A pitch target approximation model for f0 contours in Mandarin. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 2359–2362.

Yi Xu. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25:62–83.

Y. Xu. 1999. Effects of tone and focus on the formation and alignment of f0 contours - evidence from Mandarin. *Journal of Phonetics*, 27.

Yi Xu. 2004. Transmitting tone and intonation simultaneously - the parallel encoding and target approximation (PENTA) model. In *TAL-2004*, pages 215–220.