Using Conditional Random Fields to Predict Pitch Accents in Conversational Speech

Michelle L. Gregory
Linguistics Department
University at Buffalo
Buffalo, NY 14260
mgregory@buffalo.edu

Yasemin Altun
Department of Computer Science
Brown University
Providence, RI 02912
altun@cs.brown.edu

Abstract

The detection of prosodic characteristics is an important aspect of both speech synthesis and speech recognition. Correct placement of pitch accents aids in more natural sounding speech, while automatic detection of accents can contribute to better word-level recognition and better textual understanding. In this paper we investigate probabilistic, contextual, and phonological factors that influence pitch accent placement in natural, conversational speech in a sequence labeling setting. We introduce Conditional Random Fields (CRFs) to pitch accent prediction task in order to incorporate these factors efficiently in a sequence model. We demonstrate the usefulness and the incremental effect of these factors in a sequence model by performing experiments on hand labeled data from the Switchboard Corpus. Our model outperforms the baseline and previous models of pitch accent prediction on the Switchboard Corpus.

1 Introduction

The suprasegmental features of speech relay critical information in conversation. Yet, one of the major roadblocks to natural sounding speech synthesis has been the identification and implementation of prosodic characteristics. The difficulty with this task lies in the fact that prosodic cues are never absolute; they are relative to individual speakers, gender, dialect, discourse context, local context, phonological environment, and many other factors. This is especially true of pitch accent, the acoustic cues that make one word more prominent than others in an utterance. For example, a word with a fundamental frequency (f0) of 120 Hz would likely be quite prominent in a male speaker, but not for a typical female speaker. Likewise, the accent on the utterance "*Jon's leaving.*" is critical in determining whether it is the answer to the question "*Who is leaving?*" ("*JON's leaving.*") or "*What is Jon doing?*" ("*Jon's LEAVING.*"). Accurate pitch accent prediction lies in the successful combination of as many of the contextual variables as possible. Syntactic information such as part of speech has proven to be a successful predictor of accentuation (Hirschberg, 1993; Pan and Hirschberg, 2001). In general, function words are not accented, while content words are. Various measures of a word's informativeness, such as the information content (IC) of a word (Pan and McKeown, 1999) and its collocational strength in a given context (Pan and Hirschberg, 2001) have also proven to be useful models of pitch accent. However, in open topic conversational speech, accent is very unpredictable. Part of speech and the informativeness of a word do not capture all aspects of accentuation, as we see in this example taken from Switchboard, where a function word gets accented (accented words are in uppercase):

I, I have STRONG OBJECTIONS to THAT.

Accent is also influenced by aspects of rhythm and timing. The length of words, in both number of phones and normalized duration, affect its likelihood of being accented. Additionally, whether the immediately surrounding words bear pitch accent also affect the likelihood of accentuation. In other words, a word that might typically be accented may be unaccented because the surrounding words also bear pitch accent. Phrase boundaries seem to play a role in accentuation as well. The first word of intonational phrases (IP) is less likely to be accented while the last word of an IP tends be accented. In short, accented words within the same IP are not independent of each other.

Previous work on pitch accent prediction, however, neglected the dependency between labels. Different machine learning techniques, such as decision trees (Hirschberg, 1993), rule induction systems (Pan and McKeown, 1999), bagging (Sun, 2002), boosting (Sun, 2002) have been used in a scenario where the accent of each word is predicted independently. One exception to this line of research is the use of Hidden Markov Models

(HMM) for pitch accent prediction (Pan and McKeown, 1999; Conkie et al., 1999). Pan and McKeown (1999) demonstrate the effectiveness of a sequence model over a rule induction system, RIPPER, that treats each label independently by showing that HMMs outperform RIPPER when the same variables are used.

Until recently, HMMs were the predominant formalism to model label sequences. However, they have two major shortcomings. They are trained non-discriminatively using maximum likelihood estimation to model the joint probability of the observation and label sequences. Also, they require questionable independence assumptions to achieve efficient inference and learning. Therefore, variables used in Hidden Markov models of pitch accent prediction have been very limited, e.g. part of speech and frequency (Pan and McKeown, 1999). Discriminative learning methods, such as Maximum Entropy Markov Models (McCallum et al., 2000), Projection Based Markov Models (Punyakanok and Roth, 2000), Conditional Random Fields (Lafferty et al., 2001), Sequence AdaBoost (Altun et al., 2003a), Sequence Perceptron (Collins, 2002), Hidden Markov Support Vector Machines (Altun et al., 2003b) and Maximum-Margin Markov Networks (Taskar et al., 2004), overcome the limitations of HMMs. Among these methods, CRFs is the most common technique used in NLP and has been successfully applied to Part-of-Speech Tagging (Lafferty et al., 2001), Named-Entity Recognition (Collins, 2002) and shallow parsing (Sha and Pereira, 2003; McCallum, 2003).

The goal of this study is to better identify which words in a string of text will bear pitch accent. Our contribution is two-fold: employing new predictors and utilizing a discriminative model. We combine the advantages of probabilistic, syntactic, and phonological predictors with the advantages of modeling pitch accent in a sequence labeling setting using CRFs (Lafferty et al., 2001).

The rest of the paper is organized as follows: In Section 2, we introduce CRFs. Then, we describe our corpus and the variables in Section 3 and Section 4. We present the experimental setup and report results in Section 5. Finally, we discuss our results (Section 6) and conclude (Section 7).

2 Conditional Random Fields

CRFs can be considered as a generalization of logistic regression to label sequences. They define a conditional probability distribution of a label sequence $\mathbf{y}$ given an observation sequence $\mathbf{x}$. In this paper, $\mathbf{x} = (x^1, x^2, \ldots, x^n)$ denotes a sentence of length n and $\mathbf{y} = (y^1, y^2, \ldots, y^n)$ denotes the label sequence corresponding to $\mathbf{x}$. In pitch accent prediction, x^t is a word and y^t is a binary label denoting whether x^t is accented or not.

CRFs specify a linear discriminative function F parameterized by Λ over a feature representation of the observation and label sequence $\Psi(\mathbf{x}, \mathbf{y})$. The model is assumed to be stationary, thus the feature representation can be partitioned with respect to positions t in the sequence and linearly combined with respect to the *importance* of each feature ψ_k, denoted by λ_k. Then the discriminative function can be stated as in Equation 1:

$$F(\mathbf{x}, \mathbf{y}; \Lambda) = \sum_t \langle \Lambda, \Psi_t(\mathbf{x}, \mathbf{y}) \rangle \quad (1)$$

Then, the conditional probability is given by

$$p(\mathbf{y}|\mathbf{x}; \Lambda) = \frac{1}{Z(\mathbf{x}, \Lambda)} F(\mathbf{x}, \mathbf{y}; \Lambda) \quad (2)$$

where $Z(\mathbf{x}, \Lambda) = \sum_{\bar{\mathbf{y}}} F(\mathbf{x}, \bar{\mathbf{y}}; \Lambda)$ is a normalization constant which is computed by summing over all possible label sequences $\bar{\mathbf{y}}$ of the observation sequence $\mathbf{x}$.

We extract two types of features from a sequence pair:

1. Current label and information about the observation sequence, such as part-of-speech tag of a word that is within a window centered at the word currently labeled, e.g. *Is the current word pitch accented and the part-of-speech tag of the previous word=Noun?*
2. Current label and the neighbors of that label, i.e. features that capture the inter-label dependencies, e.g. *Is the current word pitch accented and the previous word not accented?*

Since CRFs condition on the observation sequence, they can efficiently employ feature representations that incorporate overlapping features, i.e. multiple interacting features or long-range dependencies of the observations, as opposed to HMMs which generate observation sequences.

In this paper, we limit ourselves to 1-order Markov model features to encode inter-label dependencies. The information used to encode the observation-label dependencies is explained in detail in Section 4.

In CRFs, the objective function is the log-loss of the model with Λ parameters with respect to a training set $\mathcal{D}$. This function is defined as the negative

sum of the conditional probabilities of each training label sequence $\mathbf{y}_i$, given the observation sequence $\mathbf{x}_i$, where $\mathcal{D} \equiv \{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, m\}$. CRFs are known to overfit, especially with noisy data if not regularized. To overcome this problem, we penalize the objective function by adding a Gaussian prior (a term proportional to the squared norm $||\Lambda||^2$) as suggested in (Johnson et al., 1999). Then the loss function is given as:

$$\begin{aligned} \mathcal{L}(\Lambda; \mathcal{D}) &= -\sum_{i}^{m} \log p(\mathbf{y}_i|\mathbf{x}_i; \Lambda) + \frac{1}{2} c||\Lambda||^2 \\ &= -\sum_{i}^{m} F(\mathbf{x}_i, \mathbf{y}_i; \Lambda) + \log Z(\mathbf{x}_i, \Lambda) \\ &+ \frac{1}{2} c||\Lambda||^2 \end{aligned} \quad (3)$$

where c is a constant.

Lafferty et al. (2001), proposed a modification of improved iterative scaling for parameter estimation in CRFs. However, gradient-based methods have often found to be more efficient for minimizing Equation 3 (Minka, 2001; Sha and Pereira, 2003). In this paper, we use the conjugate gradient descent method to optimize the above objective function. The gradients are computed as in Equation 4:

$$\begin{aligned} \nabla_\Lambda \mathcal{L} &= \sum_{i}^{m} \sum_{t} \mathbf{E}_p[\Psi_t(\mathbf{x}_i, \mathbf{y})] - \Psi_t(\mathbf{x}_i, \mathbf{y}_i) \\ &+ c\Lambda \end{aligned} \quad (4)$$

where the expectation is with respect to all possible label sequences of the observation sequence $\mathbf{x}_i$ and can be computed using the forward backward algorithm.

Given an observation sequence $\mathbf{x}$, the best label sequence is given by:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}; \hat{\Lambda}) \quad (5)$$

where $\hat{\Lambda}$ is the parameter vector that minimizes $\mathcal{L}(\Lambda; \mathcal{D})$. The best label sequence can be identified by performing the Viterbi algorithm.

3 Corpus

The data for this study were taken from the Switchboard Corpus (Godfrey et al., 1992), which consists of 2430 telephone conversations between adult speakers (approximately 2.4 million words). Participants were both male and female and represented all major dialects of American English. We used a portion of this corpus that was phonetically hand-transcribed (Greenberg et al., 1996) and segmented into speech boundaries at turn boundaries or pauses of more than 500 ms on both sides. Fragments contained seven words on average. Additionally, each word was coded for probabilistic and contextual information, such as word frequency, conditional probabilities, the rate of speech, and the canonical pronunciation (Fosler-Lussier and Morgan, 1999).

The dataset used in all analysis in this study consists of only the first hour of the database, comprised of 1,824 utterances with 13,190 words. These utterances were hand coded for pitch accent and intonational phrase brakes.

3.1 Pitch Accent Coding

The utterances were hand labeled for accents and boundaries according to the Tilt Intonational Model (Taylor, 2000). This model is characterized by a series of intonational events: accents and boundaries. Labelers were instructed to use duration, amplitude, pausing information, and changes in f0 to identify events. In general, labelers followed the basic conventions of EToBI for coding (Taylor, 2000). However, the Tilt coding scheme was simplified. Accents were coded as either major or minor (and some rare level accents) and breaks were either rising or falling. Agreement for the Tilt coding was reported at 86%. The CU coding also used a simplified EToBI coding scheme, with accent types conflated and only major breaks coded. Accent and break coding pair-wise agreement was between 85-95% between coders, with a kappa κ of 71%-74% where κ is the difference between expected agreement and actual agreement.

4 Variables

The label we were predicting was a binary distinction of accented or not. The variables we used for prediction fall into three main categories: syntactic, probabilistic variables, which include word frequency and collocation measures, and phonological variables, which capture aspects of rhythm and timing that affect accentuation.

4.1 Syntactic variables

The only syntactic category we used was a four-way classification for hand-generated part of speech (POS): Function, Noun, Verb, Other, where Other includes all adjectives and adverbs[1]. Table 1 gives the percentage of accented and unaccented items by POS.

[1]We also tested a categorization of 14 distinct part of speech classes, but the results did not improve, so we only report on the four-way classifi cation.

	Accented	Unaccented
Function	21%	79%
Verb	59%	41%
Noun	30%	70%
Other	49%	51%

Table 1: Percentage of accented and unaccented items by POS.

Variable	Definition	Example
Unigram	$\log p(w_i)$	*and, I*
Bigram	$\log p(w_i \mid w_{i-1})$	roughing *it*
Rev Bigram	$\log p(w_i \mid w_{i+1})$	*rid* of
Joint	$\log p(w_{i-1}, w_i)$	*and* I
Rev Joint	$\log p(w_i, w_{i+1})$	and *I*

Table 2: Definition of probabilistic variables.

4.2 Probabilistic variables

Following a line of research that incorporates the information content of a word as well as collocation measures (Pan and McKeown, 1999; Pan and Hirschberg, 2001) we have included a number of probabilistic variables. The probabilistic variables we used were the unigram frequency, the predictability of a word given the preceding word (bigram), the predictability of a word given the following word (reverse bigram), the joint probability of a word with the preceding (joint), and the joint probability of a word with the following word (reverse joint). Table 2 provides the definition for these, as well as high probability examples from the corpus (the emphasized word being the current target). Note all probabilistic variables were in $\log$ scale.

The values for these probabilities were obtained using the entire 2.4 million words of SWBD[2]. Table 3 presents the Spearman's rank correlation coefficient between the probabilistic measures and accent (Conover, 1980). These values indicate the strong correlation of accents to the probabilistic variables. As the probability increases, the chance of an accent decreases. Note that all values are significant at the $p < .001$ level.

We also created a combined part of speech and unigram frequency variable in order to have a variable that corresponds to the variable used in (Pan and McKeown, 1999).

[2]Our current implementation of CRF only takes categorical variables, thus for the experiments, all probabilistic variables were binned into 5 equal categories. We also tried more bins and produced similar results, so we only report on the 5-binned categories. We computed correlations between pitch accent and the original 5 variables as well as the binned variables and they are very similar.

Variables	Spearman's ρ
Unigram	-.451
Bigram	-.309
Reverse Bigram	-.383
Joint	-.207
Reverse joint	-.265

Table 3: Spearman's correlation values for the probabilistic measures.

4.3 Phonological variables

The last category of predictors, phonological variables, concern aspects of rhythm and timing of an utterance. We have two main sources for these variables: those that can be computed solely from a string of text (textual), and those that require some sort of acoustic information (acoustic). Sun (2002) demonstrated that the number of phones in a syllable, the number of syllables in a word, and the position of a word in a sentence are useful predictors of which syllables get accented. While Sun was concerned with predicting accented syllables, some of the same variables apply to word level targets as well. For our textual phonological features, we included the number of syllables in a word and the number of phones (both in citation form as well as transcribed form). Instead of position in a sentence, we used the position of the word in an utterance since the fragments do not necessarily correspond to sentences in the database we used. We also made use of the utterance length. Below is the list of our textual features:

- Number of canonical syllables
- Number of canonical phones
- Number of transcribed phones
- The length of the utterance in number of words
- The position of the word in the utterance

The main purpose of this study is to better predict which words in a string of text receive accent. So far, all of our predictors are ones easily computed from a string of text. However, we have included a few variables that affect the likelihood of a word being accented that require some acoustic data. To the best of our knowledge, these features have not been used in acoustic models of pitch accent prediction. These features include the duration of the word, speech rate, and following intonational phrase boundaries. Given the nature of the SWBD corpus, there are many disfluencies. Thus, we also

Feature	χ^2	Sig
canonical syllables	1636	$p < .001$
canonical phones	2430	$p < .001$
transcribed phones	2741	$p < .001$
utt length	80	$p < .005$
utt position	295	$p < .001$
duration	3073	$p < .001$
speech rate	101	$p < .001$
following pause	27	$p < .001$
foll filled pause	328	$p < .001$
foll IP boundary	1047	$p < .001$

Table 4: Significance of phonological features on pitch accent prediction.

included following pauses and filled pauses as predictors. Below is the list of our acoustic features:

- Log of duration in milliseconds normalized by number of canonical phones binned into 5 equal categories.
- Log Speech Rate; calculated on strings of speech bounded on either side by pauses of 300 ms or greater and binned into 5 equal categories.
- Following pause; a binary distinction of whether a word is followed by a period of silence or not.
- Following filled pause; a binary distinction of whether a word was followed by a filled pause (uh, um) or not.
- Following IP boundary

Table 4 indicates that each of these features significantly affect the presence of pitch accent. While certainly all of these variables are not independent of on another, using CRFs, one can incorporate all of these variables into the pitch accent prediction model with the advantage of making use of the dependencies among the labels.

4.4 Surrounding Information

Sun (2002) has shown that the values immediately preceding and following the target are good predictors for the value of the target. We also experimented with the effects of the surrounding values by varying the window size of the observation-label feature extraction described in Section 2. When the window size is 1, only values of the word that is labelled are incorporated in the model. When the window size is 3, the values of the previous and the following words as well as the current word are incorporated in the model. Window size 5 captures the values of the current word, the two previous words and the two following words.

5 Experiments and Results

All experiments were run using 10 fold cross-validation. We used Viterbi decoding to find the most likely sequence and report the performance in terms of label accuracy. We ran all experiments with varying window sizes ($w \in \{1, 3, 5\}$). The baseline which simply assigns the most common label, unaccented, achieves $60.53 \pm 1.50\%$.

Previous research has demonstrated that part of speech and frequency, or a combination of these two, are very reliable predictors of pitch accent. Thus, to test the worthiness of using a CRF model, the first experiment we ran was a comparison of an HMM to a CRF using just the combination of part of speech and unigram. The HMM score (referred as *HMM:POS, Unigram* in Table 5) was 68.62 ± 1.78, while the CRF model (referred as *CRF:POS, Unigram* in Table 5) performed significantly better at 72.56 ± 1.86. Note that Pan and McKeown (1999) reported 74% accuracy with their HMM model. The difference is due to the different corpora used in each case. While they also used spontaneous speech, it was a limited domain in the sense that it was speech from discharge orders from doctors at one medical facility. The SWDB corpus is open domain conversational speech.

In order to capture some aspects of the IC and collocational strength of a word, in the second experiment we ran part of speech plus all of the probabilistic variables (referred as *CRF:POS, Prob* in Table 5). The model accuracy was 73.94%, thus improved over the model using POS and unigram values by 1.38%.

In the third experiment we wanted to know if TTS applications that made use of purely textual input could be aided by the addition of timing and rhythm variables that can be gleaned from a text string. Thus, we included the textual features described in Section 4.3 in addition to the probabilistic and syntactic features (referred as *CRF:POS, Prob, Txt* in Table 5). The accuracy was improved by 1.73%.

For the final experiment, we added the acoustic variable, resulting in the use of all the variables described in Section 4 (referred as *CRF:All* in Table 5). We get about 0.5% increase in accuracy, 76.1% with a window of size $w = 1$.

Using larger windows resulted in minor increases in the performance of the model, as summarized in Table 5. Our best accuracy was 76.36% using all features in a $w = 5$ window size.

Model:Variables	$w = 1$	$w = 3$	$w = 5$
Baseline	60.53		
HMM: POS,Unigram	68.62		
CRF: POS, Unigram	72.56		
CRF: POS, Prob	73.94	74.19	74.51
CRF: POS, Prob, Txt	75.67	75.74	75.89
CRF: All	76.1	76.23	76.36

Table 5: Test accuracy of pitch accent prediction on SWDB using various variables and window sizes.

6 Discussion

Pitch accent prediction is a difficult task, in that, the number of different speakers, topics, utterance fragments and disfluent production of the SWBD corpus only increase this difficulty. The fact that 21% of the function words are accented indicates that models of pitch accent that mostly rely on part of speech and unigram frequency would not fair well with this corpus. We have presented a model of pitch accent that captures some of the other factors that influence accentuation. In addition to adding more probabilistic variables and phonological factors, we have used a sequence model that captures the interdependence of accents within a phrase.

Given the distinct natures of corpora used, it is difficult to compare these results with earlier models. However, in experiment 1 (*HMM: POS, Unigram* vs *CRF: POS, Unigram*) we have shown that a CRF model achieves a better performance than an HMM model using the same features. However, the real strength of CRFs comes from their ability to incorporate different sources of information efficiently, as is demonstrated in our experiments.

We did not test directly the probabilistic measures (or collocation measures) that have been used before for this task, namely information content (IC) (Pan and McKeown, 1999) and mutual information (Pan and Hirschberg, 2001). However, the measures we have used encompass similar information. For example, IC is only the additive inverse of our unigram measure:

$$IC(w) = -\log p(w) \tag{6}$$

Rather than using mutual information as a measure of collocational strength, we used unigram, bigram and joint probabilities. A model that includes both joint probability and the unigram probabilities of w_i and w_{i-1} is comparable to one that includes mutual information.

Just as the likelihood of a word being accented is influenced by a following silence or IP boundary, the collocational strength of the target word with the following word (captured by reverse bigram and reverse joint) is also a factor. With the use of POS, unigram, and all bigram and joint probabilities, we have shown that (a) CRFs outperform HMMs, and (b) our probabilistic variables increase accuracy from a model that include POS + unigram (73.94% compared to 72.56%).

For tasks in which pitch accent is predicted solely based on a string of text, without the addition of acoustic data, we have shown that adding aspects of rhythm and timing aids in the identification of accent targets. We used the number of words in an utterance, where in the utterance a word falls, how long in both number of syllables and number of phones all affect accentuation. The addition of these variables improved the model by nearly 2%. These results suggest that Accent prediction models that only make use of textual information could be improved with the addition of these variables.

While not trying to provide a complete model of accentuation from acoustic information, in this study we tested a few acoustic variables that have not yet been tested. The nature of the SWBD corpus allowed us to investigate the role of disfluencies and widely variable durations and speech rate on accentuation. Especially speech rate, duration and surrounding silence are good predictors of pitch accent. The addition of these predictors only slightly improved the model (about .5%). Acoustic features are very sensitive to individual speakers. In the corpus, there are many different speakers of varying ages and dialects. These variables might become more useful if one controls for individual speaker differences. To really test the usefulness of these variables, one would have to combine them with acoustic features that have been demonstrated to be good predictors of pitch accent (Sun, 2002; Conkie et al., 1999; Wightman et al., 2000).

7 Conclusion

We used CRFs with new measures of collocational strength and new phonological factors that capture aspects of rhythm and timing to model pitch accent prediction. CRFs have the theoretical advantage of incorporating all these factors in a principled and efficient way. We demonstrated that CRFs outperform HMMs also experimentally. We also demonstrated the usefulness of some new probabilistic variables and phonological variables. Our results mainly have implications for the textual prediction of accents in TTS applications, but might also be useful in automatic speech recognition tasks such as automatic transcription of multi-speaker meetings. In the near future we would like to incorporate reliable acoustic

information, controlling for individual speaker difference and also apply different discriminative sequence labeling techniques to pitch accent prediction task.

8 Acknowledgements

This work was partially funded by CAREER award #IIS 9733067 IGERT. We would also like to thank Mark Johnson for the idea of this project, Dan Jurafsky, Alan Bell, Cynthia Girand, and Jason Brenier for their helpful comments and help with the database.

References

Y. Altun, T. Hofmann, and M. Johnson. 2003a. Discriminative learning for label sequences via boosting. In *Proc. of Advances in Neural Information Processing Systems*.

Y. Altun, I. Tsochantaridis, and T. Hofmann. 2003b. Hidden markov support vector machines. In *Proc. of 20th International Conference on Machine Learning*.

M. Collins. 2002. Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. In *Proc. of Empirical Methods of Natural Language Processing*.

A. Conkie, G. Riccardi, and R. Rose. 1999. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In *Proc. of EUROSPEECH'99*.

W. J. Conover. 1980. *Practical Nonparametric Statistics*. Wiley, New York, 2nd edition.

E. Fosler-Lussier and N. Morgan. 1999. Effects of speaking rate and word frequency on conversational pronunci ations. In *Speech Communication*.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and develo pment. In *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*.

S. Greenberg, D. Ellis, and J. Hollenback. 1996. Insights into spoken language gleaned from phonetic transcripti on of the Switchboard corpus. In *Proc. of International Conference on Spoken Language Processsing*.

J. Hirschberg. 1993. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340.

M. Johnson, S. Geman, S. Canon, Z. Chi, and S. Riezler. 1999. Estimators for stochastic unification-based grammars. In *Proc. of ACL'99 Association for Computational Linguistics*.

J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of 18th International Conference on Machine Learning*.

A. McCallum, D. Freitag, and F. Pereira. 2000. Maximum Entropy Markov Models for Information Extraction and Segmentation. In *Proc. of 17th International Conference on Machine Learning*.

A. McCallum. 2003. Efficiently inducing features of Conditional Random Fields. In *Proc. of Uncertainty in Articifical Intelligence*.

T. Minka. 2001. Algorithms for maximum-likelihood logistic regression. Technical report, CMU, Department of Statistics, TR 758.

S. Pan and J. Hirschberg. 2001. Modeling local context for pitch accent prediction. In *Proc. of ACL'01, Association for Computational Linguistics*.

S. Pan and K. McKeown. 1999. Word informativeness and automatic pitch accent modeling. In *Proc. of the Joint SIGDAT Conference on EMNLP and VLC*.

V. Punyakanok and D. Roth. 2000. The use of classifiers in sequential inference. In *Proc. of Advances in Neural Information Processing Systems*.

F. Sha and F. Pereira. 2003. Shallow parsing with conditional random fields. In *Proc. of Human Language Technology*.

Xuejing Sun. 2002. Pitch accent prediction using ensemble machine learning. In *Proc. of the International Conference on Spoken Language Processing*.

B. Taskar, C. Guestrin, and D. Koller. 2004. Max-margin markov networks. In *Proc. of Advances in Neural Information Processing Systems*.

P. Taylor. 2000. Analysis and synthesis of intonation using the Tilt model. *Journal of the Acoustical Society of America*.

C. W. Wightman, A. K. Syrdal, G. Stemmer, A. Conkie, and M. Beutnagel. 2000. Perceptually Based Automatic Prosody Labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-To-Speech Synthesis. volume 2, pages 71–74.