PROSODY PREDICTION FOR SPEECH SYNTHESIS USING TRANSFORMATIONAL RULE-BASED LEARNING

Cameron S. Fordyce[1] *Mari Ostendorf*

ECE Department, Boston University, Boston, Massachusetts, USA
[1]Now with Lernout and Hauspie Speech Products, Burlington, Massachusetts, USA

ABSTRACT

Prediction of symbolic prosodic labels (pitch accents and phrase structure) is an important step in generating natural synthetic speech. This paper investigates a new automatically trainable procedure for combined accent and phrase prediction based on transformational rule-based learning. Experimental results on a radio news corpus show that accent prediction benefits from phrase structure, but not vice versa, and that TRBL outperforms simple decision tree predictors.

1. INTRODUCTION

Although most speech synthesis systems are typically judged to be intelligible in simple listening tests, the ability to understand synthetic speech degrades quickly in noisy or high cognitive load situations [1]. Furthermore, the output compares poorly with human speech in terms of naturalness, which directly impacts widespread acceptance of synthesis in spoken dialog systems. Many researchers have proposed that prosody is an important area in which intelligibility and naturalness of synthesized speech can be improved, e.g. [2]. Prosody is often predicted in two stages: first symbolic labels are assigned, and then continuous features such as duration and fundamental frequency are predicted. The aspect of prosody that this work will focus on is prediction of symbolic accent and phrase boundary labels.

Although some successful systems use handwritten rules, there is a growing interest in automatically trainable approaches since they can be easily ported to new speaking styles and task domains. To date, the most successful automatic algorithms for predicting symbolic prosodic labels are based on decision trees and/or Markov models [3, 4, 5, 6, 7]. These approaches have the disadvantage of being sensitive to training data limitations, a problem because of the high cost of prosodic labeling. This work will present a new approach based on transformational rule-based learning (TRBL) that is more robust in sparse training conditions. The TRBL algorithm produces an ordered sequence of rules that transforms unlabeled data to labeled data. The algorithm was introduced by Brill [8] for the part-of-speech tagging task. In our case, the labels indicate pitch accent and phrase boundary locations.

In the next section, we present an overview of the decision tree and TRBL prediction algorithms, which will then be compared experimentally. For the synthesis application, both methods can be thought of as automatic rule learning procedures, but with different advantages.

2. AUTOMATIC RULE LEARNING

2.1. Decision Trees

A decision tree is an ordered sequence of questions about elements of a feature vector that lead to classification of that vector. Questions may be of the form "is $x_i > T$?" or "is $x_i \in A$?". Based on the answer to the questions, which can be thought of as "if-then" rules, one traverses a tree until reaching a terminal node which has the class label assignment. The order of the questions and the parameters (e.g. T, A) are trained automatically from data by using a greedy algorithm that associates subsets of training data with each node in the tree and successively partitions the data into finer and finer subsets with the questions that most improve an overall objective function [9]. For classification, where the ultimate goal is typically minimum error, the objective function is minimum "impurity" of the class probability distributions associated with terminal nodes. Impurity may simply be classification error rate, but other indirect measures tend to be more robust and minimum entropy is used here. The size of the tree can be determined by a simple stopping criterion or by using more sophisticated techniques that look at independent data (e.g. cross validation). An inherent problem in decision tree training is that the amount of data available at nodes lower in the tree is diminished with each new split. A related problem is that data at one node in the tree is not available to other nodes in question design, which is appropriate for questions based on dependent features but not ideal if there are independent features.

2.2. TRBL

Transformational rule-based learning is a machine learning algorithm which finds an ordered sequence of rules that iteratively minimize the overall classification error. The resulting rule sequence can then be applied to labeling new data. Each rule is chosen by a greedy search over the entire corpus. The search for new rules stops when the decrease in misclassification accuracy reaches a minimum threshold. Brill first introduced this method to the natural language processing community [8] for the part-of-speech tagging task. In that task, TRBL performed on par with the contemporary stochastic methods often with much less data (64k words vs. 1M). TRBL has also been shown to have equivalent or greater success than other algorithms in such diverse language processing tasks as prepositional phrase attachment [10] and word segmentation [11].

TRBL has three basic requirements to define the system [8]: an initial-state annotator, a set of prototype rules, and a function for ranking potential rules. Sensitivity to the initial state has been reported in some though not all applications, so we investigated several starting points. The prototypes or "templates" provided to the learning algorithm are analogous to the set of allowable questions in decision tree training, except that in TRBL the questions are associated with an explicit rule *change* rather than the implicit label assignment in decision trees. The set allowable templates should be of reasonable size and complexity to limit computational costs, but also broad enough to capture important dependencies in the data. Finally, the ranking function is a score computed for each possible instantiation of a rule by calculating the improvement in classifier performance. In the applications explored here, performance is measured in terms of classification error for accent assignment and an absolute distance for phrase prediction, as discussed further in section 3.

Given the three components described above, the learning process proceeds as follows:

1. Label the data according to some base rules to create an initial state.
2. For all possible rule prototypes, all possible features, and observed[1] feature values:
   (a) Instantiate a rule from a prototype.
   (b) Apply the rule to the data.
   (c) Compare the predicted labels with the "truth" and record the performance gain (i.e. score).
3. Find the rule that maximizes the gain in accuracy (or reduction in distance).
4. Test if the new rule exceeds a threshold of minimal gain. If not, stop. Otherwise, change the candidate labels of the data according to this rule, and go to step (2).

The rule sets are similar to handwritten rules. The major difference is that while the prototypes are designed by hand, the values for the features and the order of rule application are automatically learned.

2.3. Similarities and Trade-offs

Decision tree design and TRBL are quite similar in that they are both automatic methods of learning rules for labeling data and both are designed with greedy algorithms. Both have the advantages of automatic training from data, which allows the algorithm to capture relationships that might be missed by a human expert and to learn the relative importance of rules that may vary with speaking style. However, there are important differences between decision trees and TRBL. As mentioned above, TRBL rules involve successively changing the labeling of a data sequence, i.e. placing intermediate labels on the data that provide a sort of hidden state in rule learning. TRBL can easily and inexpensively incorporate rules about other label assignments for data in a sequence. In contrast, decision trees are aimed at classifying independent vectors, though questions about local context can be incorporated by making Markov assumptions and using dynamic programming to find the most likely sequence [6]. TRBL allows for more data to be examined in the design of each rule, since the entire space of data sharing feature values is examined for a question. Compare this approach to decision trees where questions use only the data subset of the associated node. For this reason, TRBL tends to be less sensitive to data sparsity, and is better able to learn parameters associated with independent factors. One disadvantage of TRBL is that, by not associating probabilities with the label assignments, it is not as well suited for recognition applications.

3. EXPERIMENTS

3.1. Paradigm

Experiments are based on the Boston University Radio News corpus [12], training and testing with a single speaker (F2B). Recorded broadcast news stories were used as the training data (9181 words). Four different news stories read in the laboratory constitute the test data (2113 words). All the speech was hand transcribed. The corpus was labeled with part-of-speech tags using an HMM-based tagger [13]. Part-of-speech labeling errors were corrected for the test data, but not the training data.

Data were prosodically labeled using the ToBI system[14], which includes pitch accent tones, breaks, and phrase boundary tones. For this study, only presence vs. absence of pitch accents are used, and phrase boundary markers are collapsed into three categories: major, minor, and no boundary. A consistency study of human labelers transcribing this corpus found that there was 91% agreement among labelers for presence vs. absence of pitch accents [6]. Inter-annotator agreement was 93% for the location of phrase boundaries and 91% agreement for the location of phrase accents. To some extent, the inter-annotator agreement provides and upper bound on performance of the prediction algorithms. In the test data, 48% of the words and 31% of the syllables receive pitch accents, which provides a lower bound for achievable error rates. Of the test words, 20% are followed by major boundaries, 8% by minor boundaries, and 72% are not followed by a boundary.

Phrase and accent location are evaluated in comparison to a target version, which is the hand-labeled prosodic transcription of test sentences read by the target speaker in the news broadcasting style. Pitch accent prediction accuracy is the number of correctly predicted labels over the

[1] To reduce the cost of the search, values for template features are chosen from the values that occur in the error space.

total possible. Results are reported at the syllable level, since prediction is at that level to capture the phenomenon of early accent placement [15]. Phrase boundary prediction is evaluated in terms of average absolute distance from the target boundary:

$$D_{phr} = \frac{1}{N}\Sigma_{i=1}^{N}|\alpha_i - \hat{\alpha}_i|, \qquad (1)$$

where phrase boundaries are specified as $\alpha_i \in \{0 = none, 1 = minor, 2 = major\}$, $\hat{\alpha}_i$ corresponds to a predicted label, and N is the total number of boundaries. This metric penalizes errors of assigning no boundary for a major boundary (and vice versa) more than minor boundary assignment errors. The motivation for a distance is the observation that phenomena like duration lengthening and pause duration have a graded behavior [16] that is better reflected in a distance measure, and the finding that two labelers are much more likely to disagree on whether a phrase is major vs. minor than on whether there is a major boundary vs. none at all. Boundary accuracy rates are also reported (number of correct predictions over the total number of words), but with the caveat that TRBL performance is an under-estimate since the boundary prediction algorithm here was designed with distance rather than accuracy as a criterion.

3.2. Pitch Accent Location Prediction

The first experiment was the prediction of pitch accent locations with phrase boundaries known. Features used for prediction include lexical stress, vowel quality, part-of-speech labels, and hand-labeled phrase boundaries. Results for decision tree prediction and TRBL are presented in table 1. In order to provide a baseline, accuracy using a simple rule that assigns an accent to the syllable with primary lexical stress in every content word is shown. The constrained TRBL experiment uses rule templates that were equivalent to the types of questions that could be asked in the decision tree, with the exception that the decision tree includes questions about groups of categorical features and the TRBL templates considers only one or two values. For this constrained case, where TRBL is at somewhat of a disadvantage, we find that TRBL gives slightly better performance than the decision tree. Surprisingly, the difference is not significantly affected by reducing the training data (up to 2/3 reduction). The best case TRBL includes a two-feature combination ("and") rule template and performance improves. In both cases of TRBL, several different initialization rules were tried, with only insignificant differences in performance. Results reported here are based on initializing all syllables with a pitch accent, in which case the first rules learned effectively lead to the simple content-word system.

Table 1: Summary of pitch accent prediction experiments with phrase boundaries given. Accuracy is reported at the syllable level.

Algorithm	Accuracy
Content Word Rule	83.8%
Decision Tree	85.6%
Constrained TRBL	86.0%
Best case TRBL	86.8%

3.3. Phrase Boundary Prediction

Phrase boundary prediction experiments with known accent location were conducted next using part-of-speech, punctuation, and hand-labeled pitch accent location on words as features. Table 2 summarizes the results in terms of absolute accuracy and average phrase break distance. Again, the constrained TRBL case includes only rule templates that are equivalent to decision tree questions. The best case TRBL includes two-feature questions and allows questions about neighboring phrase boundary labels. The constrained algorithm does not produce significantly worse results than the best case. As for accent prediction, initialization does not have a significant impact on performance, and results reported here are based on initializing with no phrase breaks. Table 3 gives the label confusions for the best case TRBL system.

Table 2: Summary of phrase boundary prediction with accuracy rates and average absolute distance at the word level.

Algorithm	Distance	Accuracy
All no boundary	0.478	71.9%
Punctuation	0.266	76.9%
Decision Tree	0.253	84.1%
Constrained TRBL	0.239	82.3%
TRBL	0.235	82.6%

For phrase boundary prediction case, it is difficult to compare the decision tree and TRBL results, since the two systems are designed under different criteria.[2] TRBL is designed according to the distance criterion and as expected achieves better performance under that criterion and worse performance under the accuracy measure. Similarly, the decision tree is not designed to minimize distance; minimum entropy is better suited to the exact match (accuracy) criterion. A consequence of the difference in design criteria is that the decision tree, unlike the TRBL rules, never predicts minor boundaries because they are relatively infrequent.

Table 3: Confusion table for best case TRBL rules for phrase prediction at the word level.

	Truth		
Predicted	Major	Minor	None
Major	274	43	37
Minor	50	55	65
None	92	80	1416
Total	416	178	1518

[2] The objective function is not a fundamental difference between the models, since both could be used with either criterion.

Table 4: Pitch accent prediction accuracy using different boundary location input and accent prediction training.

Boundaries	Accent Rules	Accuracy
Hand-labeled	Original	86.8%
Predicted	Original	86.3%
Predicted	Retrained	86.7%

3.4. Combined Prediction

In the accent prediction experiments, where known boundaries are a possible prediction variable, the TRBL algorithm does in fact choose rules based on phrase boundary location. However, the converse is not true: accent location is not used in phrase boundary prediction. This suggests that a fully automatic system should first predict phrase boundaries and then accents. In this configuration, accuracy of boundary prediction is not affected, but accent prediction will be based on less reliable boundary features.

To assess the impact of phrase prediction, the best case rules were used with the predicted boundaries instead of the hand-labeled boundaries. Then, the pitch accent rules were retrained with the predicted boundary information, by running the boundary prediction rules on the training corpus. Table 4 shows the accuracy for the pitch accent prediction. Using the predicted rather than known boundaries degrades performance, as one might expect, but most of the loss is regained by retraining.

4. DISCUSSION

In summary, we have introduced a new approach to symbolic prosodic label prediction based on transformational rule-based learning. Experiments on phrase and accent prediction with a radio news corpus show that TRBL gives a small improvement over simple decision tree predictors, despite a more simplified approach to set membership rule design. In addition, the experiments showed that accent prediction benefits from phrase structure, but not vice versa. The use of average absolute distance is proposed as a new metric for design and evaluation of phrase prediction, which is motivated by the graded acoustic cues observed for different phrase boundaries. The metric can be modified to reflect changes in our understanding of different levels of phrase boundaries.

A surprising result was that performance differences between TRBL and decision trees did not seem to be sensitive to the amount of training data. This may be due to the nature of the corpus, since radio news is highly accented, and experiments on other data types are underway. Another question is why TRBL did not choose questions about neighboring prosodic labels, which is shown to be useful in decision tree work [5, 6]. We conjecture that this is a consequence of using left-to-right processing in label transformation, and that further improvements in performance can be obtained by evaluating different strategies.

Acknowledgments: This research was supported by the US National Science Foundation, grant IRI-5528990.

5. REFERENCES

1. Delogu, C., Paoloni, A. *et al.* (1996). "Spectral Analysis of Synthetic Speech and Natural Speech with Noise over the Telephone Line." *Proc. Inter. Conf. Spoken Language Processing,* 2, 1231-1234.
2. Boogaart, T., and Silverman, K. (1992). "Evaluating the Overall Comprehensibility of Speech Synthesizers." *Proc. Inter. Conf. Spoken Language Processing,* 2, 1207-1210.
3. Hirschberg, J. (1993). "Pitch Accent in Context: Predicting Prominence from text." *Artificial Intelligence,* 63, 305-340.
4. Wang, M. Q. and Hirschberg, J. (1992). "Automatic Classification of Intonational Phrase Boundaries." *Computer Speech and Language,* 6, 175-196.
5. Ostendorf, M., and Veilleux, N. (1994). "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location." *Computational Linguistics,* 20(1), 27-52.
6. Ross, K. and Ostendorf, M. (1996). "Prediction of Abstract Prosodic Labels for Speech Synthesis." *Computer Speech and Language,* 10, 155-185.
7. Black, A., and Taylor, P. (1998). "Assigning Phrase Breaks from Part-of-Speech Sequences." *Computer Speech and Language,* 12, 99-117.
8. Brill, E. (1995). "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging." *Computational Linguistics,* 21(4), 543-565.
9. Breiman, L. Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.
10. Brill, E. and Resnick, P.(1994). "A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation." *Proc. 16th Inter. Conf. Computational Linguistics.*
11. Palmer, D. (1997). "A Trainable Rule-Based Algorithm for Word Segmentation." *Proc. Annual Meeting Association for Computational Linguistics*, 321-328.
12. Ostendorf, M., Price, P., and Shattuck-Hufnagel, S. (1994). "*The Boston University Radio News Corpus.*" Boston University Technical Report ECE-95-001.
13. Meteer, M., Schwartz R., and Weischedel, R. (1991). "POST: Using probabilities in language processing." *Proc. Inter. Joint Conf. Artificial Intelligence.*
14. Silverman, K. *et al.* (1992). "TOBI: A standard for labeling prosody," *Proc. Inter. Conf. Spoken Language Processing,* 2, 867-870.
15. Shattuck-Hufnagel, S., Ostendorf, M., and Ross, K. (1994). "Stress shift and early pitch accent placement in lexical items in American English." *J. Phonetics,* 22, 357-388.
16. Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. (1992) "Segmental Durations in the Vicinity of Prosodic Phrase Boundaries." *J. Acoust. Soc. Am.,* 91(3), 1707-1717.