Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model

Xuejing Sun[1] *and Ted H. Applebaum*[2]

[1]Northwestern University, 2299 N. Campus Dr., Evanston, IL 60208, USA
[2]Panasonic Speech Technology Laboratory 3888 State St. Santa Barbara CA, 93105, USA
sunxj@northwestern.edu, ta@stl.research.panasonic.com

Abstract

In the current study, we propose and evaluate a new method for automatic intonational phrase break prediction based on sequences of parts-of-speech and word junctures. The proposed method uses decision trees to estimate the probability of a word juncture type (break or non-break) given a finite length window of part-of-speech values, and uses an n-gram to model the word juncture sequence. Trained on an 8,000 word database, our algorithm predicted breaks with F=77% and non-breaks with F=93%, which represents a significant improvement over the commonly used approach, which uses decision trees alone.

1. Introduction

Intonational phrase break prediction is an important step in text-to-speech synthesis, leading to prediction of prosodic boundaries. The information borne in phrase breaks is crucial to both intonation and duration modeling.

In the past, rule-based and stochastic approaches to phrase break prediction have been proposed. The rule-based approach is predicated on the correlation between prosodic boundaries and syntactic structure, and requires sets of heuristic rules written by linguistic experts. The stochastic approach to phrase break prediction, based on decision trees or Markov models, requires labeled training data but is less dependent on the art of heuristic rule writing.

There have been several studies of stochastic phrase break prediction for text-to-speech systems [5][6][8][10], which employ algorithms such as classification and regression trees (CART) [1], Markov models, etc. The features that have previously been used in phrase-break prediction include part-of-speech (POS), pitch accent, syntactic structure, duration, etc. It has been argued (e.g. in [8]) that some of these features, such as pitch accent, would not be available to a text-to-speech system at the time of phrase-break prediction. However, POS has proved to be an effective yet easily derived feature to predict phrase breaks.

In the present study, we propose a new method to predict intonational phrase breaks, which combines a decision tree and an n-gram model. Using a POS window as input feature, we employ a decision tree to obtain the probability distribution of word juncture type, i.e., break or non-break, at each location, and build an n-gram model upon the sequence of word junctures. The optimal word juncture sequence is determined by the Viterbi algorithm. Note that the n-gram model describes *a priori* probability of a word juncture sequence, which has also been used in Taylor and Black [8]. This approach reflects an assumption that a particular juncture type, i.e., break or non-break, is determined not only by POS but also by adjacent junctures.

2. The basic model

Our approach to predicting intonational phrase breaks from a sequence of part-of-speech tags is similar to the problem of HMM based part-of-speech tagging [3], in that we want to find the sequence of junctures $j_{1,n}$ which maximizes the conditional probability:

$$P(j_{1,n} \mid C_{1,n})$$

where $C_{1,n}$ denotes a POS window (a fixed-length sequence of POS tags).
Thus, we define:

$$J(C_{1,n}) = \arg\max_{j_{1,n}} \frac{P(j_{1,n}, C_{1,n})}{P(C_{1,n})} \qquad (1)$$

$$= \arg\max_{j_{1,n}} P(j_{1,n}, C_{1,n}) \qquad (2)$$

We ignore $P(C_{1,n})$ here as it is constant for all $j_{1,n}$.
Using Bayes' rule, we have:

$$P(j_{1,n}, C_{1,n}) = P(C_n \mid j_{1,n}, C_{1,n-1}) P(j_n \mid j_{1,n-1}, C_{1,n-1}) P(j_{1,n-1}, C_{1,n-1}) \qquad (3)$$

$$= P(j_1, C_1) \prod_{i=2}^{n} P(C_i \mid j_{1,i}, C_{1,i-1}) P(j_i \mid j_{1,i-1}, C_{1,i-1}) \qquad (4)$$

$$= \prod_{i=1}^{n} P(C_i \mid j_{1,i}, C_{1,i-1}) P(j_i \mid j_{1,i-1}, C_{1,i-1}) \qquad (5)$$

(We define the terms $P(j_{1,0}, C_{1,0}) = 1.0$, $j_i = j_{i,i}$, and $C_i = C_{i,i}$ in order to simplify our notation.)
Thus, with the Markov assumption, we can have:

$$P(C_i \mid j_{1,i}, C_{1,i-1}) = P(C_i \mid j_{1,i}) \qquad (6)$$

$$P(j_i \mid j_{1,i-1}, C_{1,i-1}) = P(j_i \mid j_{1,i-1}) \qquad (7)$$

where we assume the probability of POS window C_i depends on a juncture sequence, and the probability of a given juncture depends only on the preceding junctures. To simplify the problem, we make further assumptions: the probability of a POS window C_i depends only on its current juncture j_i, and the probability of a juncture only depends on the *k* preceding junctures. This is known as Markov Model of k-th order. Therefore, we have:

$$P(j_{1,n}, C_{1,n}) = \prod_{i=1}^{n} P(C_i \mid j_i) P(j_i \mid j_{i-k} \ldots j_{i-1}) \tag{8}$$

By using Bayes' rule again, we have

$$P(C_i \mid j_i) = \frac{P(C_i) P(j_i \mid C_i)}{P(j_i)} \tag{9}$$

Here again, we ignore the term $P(C_i)$, and the final formula is:

$$P(j_{1,n}, C_{1,n}) = \prod_{i=1}^{n} \frac{P(j_i \mid C_i)}{P(j_i)} P(j_i \mid j_{i-k} \ldots j_{i-1}) \tag{10}$$

Where $P(j_i \mid C_i)$ can be estimated by a maximum likelihood procedure (as in [8]) or via a decision tree derived by the CART training procedure [1] (as in the current work). $P(j_i)$ can be estimated by counting the frequency of occurrence of each juncture type. $P(j_i \mid j_{i-k} \ldots j_{i-1})$ is the n-gram model of juncture sequences. The optimal sequence of word juncture values may be determined from these probabilities via the Viterbi algorithm. The current algorithm is similar to that of Taylor and Black [8]. However, it differs in the method of estimating the probability of a juncture type given the part-of-speech window, i.e., $P(j_i \mid C_i)$. Taylor and Black [8] relied on frequency counting statistics, whereas we use decision trees, which give better performance and enable the incorporation of information from diverse features.

3. Corpus and evaluation

Training and testing data were taken from Boston University Radio Speech Corpus [2], speaker F2B. The database, which consists of 34 news stories read by a female professional announcer, is labeled using the ToBI [7] labeling system, as follows.

The phrase break label files contain break indices ranging from 0 to 4, where the larger number represents a higher degree of de-coupling between each pair of words. Specifically, break index 4 refers to intonational phrase boundary, also major break; 3 refers to intermediate phrase boundary, also minor break; 0-2 are generally regarded as non-breaks. In the present study, we only treat intonational phrase break, that is, regarding index 4 as a word break (BB) and all other numbers as non-break (NB).

The part-of-speech (POS) label files contain words and corresponding POS tags. The tagset includes 36 unique POS tags, which are basically the same as Penn Treebank convention. For punctuation comma, we assign a new POS tag "COMMA". Similarly, period, question mark, exclamation mark and colon (".?!:") are assigned the new POS tag "SENT". The corresponding juncture label is "DM", which stands for dummy. The purpose of the DM juncture label is to differentiate punctuation from regular POS tags, and give the n-gram model more information.

Excluding the files with mislabeled POS or break index, in total we generate a database from the corpus with 8972 words, in which there are 2038 intonational phrase breaks. The data set is split into training and testing sets with approximately a 9:1 ratio.

An n-gram model is trained to predict $P(j_i)$, the unigram probabilities of Break or Non-Break at each word juncture, as well as $P(j_i \mid j_{i-k} \ldots j_{i-1})$, the conditional probabilities of each juncture type given the juncture types for a fixed number of previous word junctures. Training data for the n-gram model are derived from the ToBI break label files. ToBI break labels are mapped to Break or Non-Break by a threshold as described above. The n-gram model is created with the CMU Statistical Language Modeling Toolkit [4]. $P(j_i \mid j_{i-k} \ldots j_{i-1})$ forms the transition probabilities of the network shown in Eq. 10. A binary decision tree is trained to predict $P(j_i \mid C_i)$, the conditional probabilities of each juncture type, given the part-of-speech tags for a fixed number of consecutive words (i.e., a "POS window"). Training data for the decision tree are derived from the join of the ToBI break label files and the POS label files, augmented with labels for punctuation as described above. The decision tree is generated by "Wagon", an implementation of the CART algorithm, from the Edinburgh Speech Tool Library [9]. The values of $P(j_i \mid C_i)$ divided by $P(j_i)$ are the observation probabilities of the Markov model shown in Eq. 10. The best sequence of word junctures is found by a Viterbi search of this network.

Although a perceptual evaluation would be ideal for a phrase break prediction system, we use objective measurements as performance criteria in order to compare with previous approaches. Similar to Koehn *et al.*[5], we compute precision, recall, and F values. These accuracy measures are defined as follows:

Precision = Number of Predicted Correct / Number of Predicted

Recall = Number of Predicted Correct / Number of True Correct

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} \tag{11}$$

where P is precision, R is recall, and $\alpha = 0.5$ is commonly used.

4. Results

In this section, we present results from a series experiments on intonational phrase break prediction. In the first experiment, only CART was used, which utilizes part-of-speech information to predict the current juncture type. We examined POS window length from 1 to 10. For a given POS window length, we tried two configurations: (1) One POS tag after the juncture, with others before the juncture; (2) Two POS tags after the juncture, with others before the juncture. In general, configuration (1) gave better results than configuration (2). The best results were obtained with POS window length of three or four.

In Table 1, we list intonational phrase break prediction results obtained from CART decision trees alone. The POS window length was four, with three POS before the juncture, and one after the juncture. We regard this as our baseline system for the purpose of comparison.

Table 1: Baseline intonational phrase break prediction performance obtained from CART alone using a full POS tagset (36 POS tags plus 2 tags for punctuation).

Juncture	**Accuracy Measure**		
	Precision (%)	**Recall (%)**	**F (%)**
Break	79.12	68.246	73.28
Non-break	90.48	94.370	92.38

In the second experiment, we used the same configuration as that in our baseline system, except for a smaller POS tagset. Taylor and Black [8] shows that the tagset size can affect the prediction performance, and achieved best results with a tagset of 23 POS tags excluding punctuation. We assume the same trend holds with our approach. Thus, we reduced the POS tagset in the same manner, and repeated the experiment using CART. This yielded slightly better results in terms of F measurement, which are shown in Table 2.

Table 2: Intonational phrase break prediction performance obtained from CART alone using a smaller tagset (23 POS tags plus 2 tags for punctuation).

Juncture	**Accuracy Measure**		
	Precision (%)	**Recall (%)**	**F (%)**
Break	77.00	72.986	74.94
Non-break	91.69	93.185	92.43

Finally, we combined the results from CART decision tree with an n-gram word juncture sequence model as described in Section 2. We investigated n-gram models with length (n) ranging from 2 to 6. None of the n-gram models were smoothed. We obtained best results with a trigram model (n=3). The results are shown in Table 3. Compared with the CART-only results in Table 2, this approach represents a significant improvement.

Table 3: Intonational phrase break prediction performance obtained from CART plus trigram juncture sequence model, using small tagset.

Juncture	**Accuracy Measure**		
	Precision (%)	**Recall (%)**	**F (%)**
Break	79.50	75.36	77.37
Non-break	92.42	93.93	93.17

5. Discussion

From the experimental results above, we can see that using an n-gram to model word juncture sequences can improve phrase-break prediction performance significantly. This shows that the information contained in the word juncture sequence is indeed a useful feature. In addition to the n-gram model, we also find that an appropriate size of the POS tagset and POS window configuration are very important in phrase break prediction.

It is not easy to directly compare the current results with other approaches as different database size and performance measurement have been used in previous studies. Nevertheless, a comparable study was conducted by Koehn *et al.* [5], which follows Wang and Hirschberg [10] in using the CART-only approach to predict intonational phrase break. Besides POS, they also used other features, such as pitch accent and detailed syntactic structure. Koehn et al. [5] presented results for different training sets, and showed a significant improvement with increasing training set size. Their training set which is closest in size to ours (8000 words) contained 10,000 words. For this training set their phrase-break prediction result is F=65.5% (precision 86.6%, recall 52.7%), which is lower than our value of F=77.4% as shown in Table 3. Interestingly it is also lower, in terms of F measurement, than our results using CART alone, as shown in Tables 1 and 2. This discrepancy could possibly be due to their smaller POS tagset. (Their exact POS tagset was not listed in the paper, however, we assume it has the same tagset as that in Wang and Hirschberg [10], which uses 7 POS categories as shown in their paper.) Their POS window includes four POS tags: two before the juncture and two after. The current system also adopted a four POS tag window, but with three POS tags before the juncture, and one after. This demonstrates the importance of POS tagset size and POS window configuration.

Our approach is similar to Taylor and Black [8] in several aspects. Both systems use a Markov model approach in general, and use POS as input feature and an n-gram model on word juncture sequence. However, Taylor and Black [8] derives $P(j_i \mid C_i)$ using maximum likelihood estimation, which is less robust to sparse data problems than are decision trees. Furthermore, using decision trees to estimate the probability distribution is more flexible than frequency counting, as decision trees offer a uniform way to incorporate a wide variety of diverse features.

Note that current system employs a very simple feature set. However, we do believe that a larger and carefully constructed feature vector is crucial to gain higher performance. To select a certain feature we should consider its computational cost and whether it is readily available at this point in a text-to-speech system. Detailed syntactic features have shown to be useful with CART-only approach, where the F measurement could be improved by 1.8% [5]. The trade-off is the complexity and computational load required to include a full syntactic parser. Therefore, future work is needed to study what is the best feature set in phrase break prediction systems.

The implementation of the current algorithm is straightforward. It has already been built into a data-driven prosody generation system. The current system is also very flexible in that: (1) It can be extended to include other features in addition to POS, since the decision tree algorithm readily combines diverse features; (2) It can be extended to predict multi-levels of juncture types. In this work we predict only two levels: break vs. non-break. However ToBI labeled data is classified into five levels of word juncture. To include more levels, such as minor break, we can add another symbol in the vocabulary and rebuild the decision tree and n-gram model.

6. Conclusions

We have presented a new algorithm for intonational phrase break prediction. This stochastic approach combines a decision tree and an n-gram model. It builds a binary decision tree upon part-of-speech tags, and an n-gram model on the word juncture sequence. It is trained on a database containing about 8,000 words, and produces predicted breaks with F=77% and non-breaks with F=93%. Compared with previous approaches to phrase break prediction, our system is efficient, yet obtains comparable or better performance.

7. Acknowlegement

We would like to thank Dr. Helmut Schmid for his assistance during this work.

8. References

[1] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, P. J. (1984). *Classification and Regression Trees.* The Wadsworth Statistics/Probability Series, Wadsworth and Brooks, 1984.

[2] Boston University Radio Speech Corpus, Linguistic Data Consortium Catalog Number LDC96S36. Available at morph.ldc.upenn.edu/Catalog/LDC96S36.html

[3] Charniak, E., Hendrickson, C., Jacobson, N., and Perkowitz, M. "Equations for part-of-speech tagging", *Proceedings of the Eleventh National Conference on Artificial Intelligence*, Menlo Park: AAAI Press/MIT Press pp. 784-789, 1993.

[4] Clarkson, P., and Rosenfeld, R. "Statistical Language Modeling Using the CMU-Cambridge Toolkit", *Proceedings of Eurospeech,* Vol 5, pp. 2707-2710, 1997.

[5] Koehn, P., Abney, S., Hirschberg, J., and Collins, M. "Improving Intonational Phrasing with Syntactic Information", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* Vol 3, pp. 1289-1290, Istanbul, June, 2000.

[6] Ostendorf, M. and Veilleux, N. "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Locations". *Computational Linguistics*, 20 (1), pp. 27-54, 1994.

[7] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.,. "ToBI: A standard for labeling prosody," *Proceedings of the International Conference on Spoken Language Processing*, pp. 867-870, 1992.

[8] Taylor, P., and Black, A. "Assigning Phrase Breaks from Part of Speech Sequences", *Computer Speech and Language*. Vol. 12, pp. 99-117, 1998.

[9] Taylor, P., Black, A., and Caley, R. *Introduction to the Edinburgh Speech Tools*, 1999. Currently available at http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/.

[10] Wang, M. Q. and Hirschberg, J. "Automatic classification of intonational phrasing boundaries". Computer Speech and Language, 6 (2), pp. 175 - 196, 1992.