Predicting Prosodic Phrasing Using Linguistic Features

Tae-Jin Yoon

Department of Linguistics
University of Illinois at Urbana-Champaign, U.S.A.
tyoon@uiuc.edu

Abstract

The prosodic structure of speech is based on complex interaction within and between several different levels of linguistic, and paralinguistic organization, and is expressed in the modulation of F0, intensity, duration, and voice quality, as well as the occurrence of pauses. Even though leading theories of prosody maintain that prosody is shaped through the interaction of grammatical factors from phonology, syntax, semantics, and pragmatics [1][2][3][4], there is no consensus on how to model their interaction. I provide a new probabilistic model of the mapping between prosody and phonology, syntax, and argument structure. The model encodes phonological features, shallow syntactic constituent structure, and basic argument structure. A machine learning experiment using these features to predict prosodic phrase boundaries achieves more than 92% accuracy in predicting prosodic boundary location: 86.10% precision and recall in predicting boundary locations and 94.61% in predicting locations where no boundary is present. An experiment for predicting the strength of prosodic boundaries achieve 88.06% accuracy. This study sheds light on the relationship between prosodic phrase structure and other grammatical structures.

1. Introduction

Leading theories of prosody maintain that prosody is shaped through the interaction of grammatical factors from phonology, syntax, semantics, and pragmatics [1][2][3][4]. However, there is no consensus on how to model their interaction (cf. [4]). Proposals have been made that the prosody interface is governed by mapping rules [1], through the interaction of constraints [4], or by the representation of discourse structure and surface syntactic structure [6], and that the mapping may be probabilistic [7][8]. While it is widely accepted that syntactic and prosodic structures are not isomorphic [9], it is also often noted that the two structures are too highly correlated for their relationship to be ignored. Proponenents of rule- or constraint-based mapping (e.g., [1][4][6]) maintain that prosodic constituents are constrained within syntactic constituents, with exceptions. Proponents of probabilistic mapping (e.g., [7][8]) propose boundary prediction based on n-gram part of speech tagging. Though these models correctly predict 86-89% of prosodic boundaries, they do not directly address the effect of syntactic constituency on prosodic boundaries. Recent probabilistic models (e.g., [10][11]) make use of full syntactic parsing, but since automatic syntactic parsing is overall not very accurate (cf. [8]), despite progress in parsing technology, the practical success of such models is limited.

I provide a new probabilistic model of the mapping between prosody and phonological, syntactic, and semantic features. The model encodes phonological features, shallow syntactic constituent structure, argument structure, and named entity tags. A machine learning experiment using these features to predict prosodic phrase boundaries achieves more than 92% accuracy in predicting prosodic boundary location, and 88.06% accuracy in predicting the strength of the prosodic boundaries. This model outperforms all published models in accuracy. This study sheds light on the relationship between prosodic phrase structure and other grammatical structures. It provides a simple algorithm for modeling the interface between distinct grammatical components, and can identify how much each linguistic factor contributes to the occurrence of prosodic phrase boundaries. The study also shows that the inclusion of linguistic information in modeling prosodic events achieves the best accuracy.

2. Prosodic labels and Corpus

2.1. ToBI: Prosodic Annotation System

This study adopts the model of prosodic phrasing put forth in the ToBI (Tones and Break Indices; [13]) labeling system, based on the Beckman-Pierrehumbert Autosegmental-Metrical (AM) theory of prosodic structure [14][2][15]. Two kinds of prosodic information are encoded: 1) tonal information and 2) information on the degree of juncture as defined in the break index. The tonal inventory in the ToBI system consists of pitch accents (H*, L*, L*+H, L+H*, H+!H*), downstepped pitch accents (!H*, !H-), and phrasal tones of intermediate phrases (L-, H-) and of intonational phrases (%H, L%, H%). The ToBI model has certain advantages over competing models such as Prosodic Phonology ([1]), in (i) defining prosodic categories in terms of tone and break index features, without explicit reference to other grammatical structures such as syntax, and (ii) the ToBI system is flexible enough to serve as an interface to other levels of linguistic encoding, such as pragmatics, as exemplified by [6][16].

2.2. Corpus: Boston Radio News Corpus

The corpus used for this work was drawn from a subset of recorded FM public radio news broadcasts produced in Boston, spoken by professional radio announcers [12]. The subset of this radio news corpus, the 'labnews portion', contains multiple renditions of four news stories. The stories are originally written for broadcast but recorded by 6 (3 male and 3 female) professional radio news speakers in a laboratory setting. The script consists of about 114 sentences, with an average word count of 18. The number of sentences used for the experiment is 583. The number of word tokens is 10,548. The duration of the speech corpus is approximately one hour. The speech files are also annotated with ToBI labels.

3. Feature Extraction

Existing work shows that prosodic phrasing is affected by syntactic structure [1][6], argument structure [17], information structure [6], phonological structure [1][17], and even prosodic structure itself [18], among other linguistic factors. In the research described here, features from syntactic structure, argument structure and phonological structure, among others, are extracted. Other aspects of prosodic structure, such as the presence of Pitch Accent, may influence the location and type of prosodic phrase boundary, but such inter-prosody effects are not considered in the present study in order to facilitate comparison with prior studies that do not consider such effects.

For the phonological features, the number of phones of each word, the number of syllables of each word, and the position of primary stress within each word are extracted.

For the syntactic features, part of speech and shallow syntactic chunking are automatically extracted using the shallow syntactic parser developed by the Inductive Linguistic Knowledge (ILK) group of the University of Tilburg[1]. The syntactic chunks are non-overlapping and non-embedded syntactic constituents, and are in a way similar to the flattened syntactic structure proposed to be used for the mapping between syntactic constituents and prosodic phrasing (cf. [17]).

For the semantic features, argument structure tags such as subject, object, and predicate are automatically extracted using the shallow syntactic parser mentioned above. Argument structure features aid in categorizing the shallow syntactic chunks into their relevant grammatical roles. The argument structure is also helpful in identifying parenthetical phrases, which are acknowledged to be an important factor in grouping of prosodic phrasing, and cause errors quite often in full syntactic parsing. Named entities such as person, location, and organization are automatically tagged by using NEPackages developed by the UIUC Cognitive Computing Group[2]. Even though shallow syntactic tagging achieves better accuracy over full syntactic tagging, it is still error-prone. Named entity tagging is employed to amend errors induced by shallow syntactic tagging.

Table 1 is an example of extracted features of a sample sentence *That year Thomas Maffy, now president of the Massachusetts Bar Association, was Hennessy's law clerk.*. Note that any errors in parsing are not corrected, and dummy symbols, though not shown in the feature matrix, are used for empty features. At the end of each sentence # was inserted into the transcription as a marker of sentence boundary.

4. Machine Learning Algorithm

Machine learning can be viewed as the extraction of generalizations over a body of input data. Memory-based learning (MBL) is used for the experiment of predicting prosodic phrasing. MBL is a machine learning algorithm that classifies unseen instances based on similarity to the instances stored in the memory, and is implemented in TiMBL [19]. The MBL system contains two components: 1) a learning component which is memory-based, and 2) a performance component which is similarity-based. For example, given a new test instance X, MBL compares X to an instance Y stored in the memory, and measures the distance between X and Y. After updating the top K of its nearest neighbors, MBL takes the majority class of the K nearest neighbors as the class of X. For the current

Table 1: (1) Word, (2) Position of the word from the end of the sentence (in sentence-reverse order), (3) Number of syllables in the word, (4) Number of phones in the word, (5) Position of the primary stress within the word, (6) Part of Speech of the word, (7) The type of syntactic phrase containing the word, (8) Position of the word from the end the syntactic phrase (phrase-reverse order), (9) the Grammatical Relation of the word, (10) the type of Named Entities containing the word, (11) Position of the Named Entities which the word belongs to.

1	2	3	4	5	6	7	8	9	10	11
That	15	1	3	1	D	NP	4	Subj		
year	14	1	3	1	N	NP	3	Subj		
Thomas	13	2	5	1	N	NP	2	Sub	PER	2
Maffy	12	2	4	1	N	NP	1	Sub	PER	1
now	11	1	2	1	Av	AVP	1			
president	10	3	9	1	N	NP	1			
of	9	1	2	1	P	P	1			
the	8	1	2	1	D	NP	4			
Massachusetts	7	4	10	3	N	NP	3		ORG	3
Bar	6	1	3	1	N	NP	2		ORG	2
Association	5	5	9	4	N	NP	1		ORG	1
was	4	1	3	1	V	VP	1	Prd		
Hennessy's	3	3	7	1	N	NP	3	NPrd	PER	1
law	2	1	2	1	N	NP	2	NPrd		
clerk	1	1	5	1	N	NP	1	NPrd		
#	#	#	#	#	#	#	#	#	#	#

experiment, $K = 1$ is adopted, and a weighted distance metric is used to calculate the similarity between X and Y, as in 1:

$$\Delta(X, Y) = \sum_i w_i \delta(x_i, y_i) \tag{1}$$

The weight is calculated as in 2:

$$w_i = \frac{H(C) - \sum_{v \in V_i} P(v) \times H(C|v)}{H(v)} \tag{2}$$

where $H(C)$ is the entropy (or uncertainty) of the class labels, defined as $H(C) = -\sum_{c \in C} P(c) \log_2 P(c)$, and $H(v)$ is the entropy of a set of feature values, defined as $H(v) = -\sum_{v \in V_i} P(v) \log_2 P(v)$. The weight, called *Gain Ratio*, is *Information Gain* divided by *Split Info* [19]. Information Gain measures how much information each feature contributes to the knowledge of the correct class labels, and Split Info controls the undesirable effect of overestimation which some features that have large numbers of values may induce. The distance metric is calculated as in 3:

$$\delta(x_i, y_i) = \begin{cases} 0 & if \quad x_i = y_i \\ 1 & if \quad x_i \neq y_i \end{cases} \tag{3}$$

where the distance for categorical variables is measured by counting the number of mismatching feature-values in both patterns, i.e., x_i and y_i. Thus, MBL can be viewed as error-induced or demotion-based learning.

5. Results

The performance of machine learning is affected by the material the learning mechanism is trained on. Thus, two issues of performance are important in evaluating the results of classification. The first is how well the learning algorithm generalizes

[1] $http://ilk.kub.nl$
[2] $http://l2r.cs.uiuc.edu$

over the training data set. The output of the machine learning algorithm can be compared to a baseline, i.e., a chance level performance. The baseline for my experiment of predicting prosodic phrasing is 73%. The second is how well the learning algorithm will perform on an unseen data set. For this purpose, 90% of the data set is used for training, and 10% of the data set is held to be used for testing.

5.1. Presence or Absence of Boundary Tone

Contextual information is used that encodes the listed features (cf. Table 1) of one word preceding, and one word following, the target word. Table 2 presents the confusion matrix of classification results for predicting the presence or absence of prosodic boundary tone. The overall accuracy, i.e., the number of correctly classified class labels divided by the total number of class labels, is 92.23%.

Table 2: Confusion matrix of presence or absence of Boundary Tone in the context of one word preceding and one word following, the target word: Overall accuracy is 92.23%. BT stands for Boundary Tone. The data are grouped according to the observed boundary tones (columns) and the predicted boundary tones (rows).

	Observ. BT	Observ. No BT
Pred. BT	**254**	41
Pred. No BT	41	**719**

Table 3 shows values of standard evaluation metrics: precision, recall, and F-value. Precision is the number of correctly predicted class labels divided by the total number of predicted class labels. Recall is the number of correctly predicted class labels divided by the total number of class labels identified as a gold standard. F-measure is the harmonic measure of precision and recall, defined as $F = \frac{2PR}{P+R}$.

Table 3: Evaluation of Presence or absence of Boundary Tones in the context of +/- 1

class	Precision	Recall	F-Score
Boundary	86.10%	86.10%	86.10%
No Boundary	94.61%	95.61%	94.61%

5.2. Strength of Prosodic Phrase Boundary

Only features of the target word where a prosodic event is observed are used. The accuracy drops when the contextual information is used. The overall accuracy of predicting the strength of the prosodic phrase boundary is 88.06%. The confusion matrix in Table 4 and the results of evaluation metrics in Table 5 reveal that ip (intermediate phrase) prediction is quite difficult to make, compared to the prediction of IP (Intonational Phrase). As is in Table 2, labels in the columns are observed phrasal tones, and labels on the rows are predicted phrasal tones.

Table 6 summarizes results from the experiment reported in [10] for prosodic strength prediction, and is shown for the purpose of comparison. The goal of [10] is to predict break indices 3 and 4, which corresponds to the prediction of ip vs. IP boundaries in the present study.

Table 4: Confusion matrix of strength of boundary tone: Overall accuracy is 88.06%

	Observ. ip	Observ. IP	Observ. No BT
Pred. ip	**29**	29	46
Pred. IP	14	**164**	11
Pred. No BT	21	12	**730**

Table 5: Evaluation of the strength of boundary tones

class	Precision	Recall	F-Score
ip	45.31%	27.88%	34.54%
IP	80.00%	86.77%	83.24%
No Boundary	92.76%	95.68%	94.19%

6. Discussion

Predicting two levels of prosodic phrase boundary from the linguistic features in this study is less accurate than simply predicting the presence or absence of these prosodic events. Nevertheless, it should be noted that the results obtained in these experiments are better than most prior studies using the same corpus or similar corpora. Table 7 shows comparison results of various learning algorithms reported in [9]. The features used in [9] are the output of a full syntactic parser.

Given the similar results across different machine learning algorithms, it is the set of features rather than the choice of a particular algorithm that counts for the better performance.

Table 8 is an example of the sentence *That year Thomas Maffy, now president of Massachusetts Bar Association, was Hennessy's law clerk.*. Each word in the sentence is aligned with the observed prosodic label and the labels predicted from the machine learning experiments reported here. In comparison, the last column in Table 8 lists the prosodic boundary labels produced by the Festival system of Text-To-Speech (TTS) synthesis [8].

7. Conclusion

This paper presents results from a machine learning experiment on the prediction of prosodic phrasing based on linguistically motivated features. This study sheds light on the relationship between prosodic phrase structure and other grammatical structures. It provides a simple probabilistic learning algorithm for modeling the interface between prosody and other components of grammar. In future work, this approach can identify how the combined linguistic factors condition the occurrence of

Table 6: Results of predicting break indices 3 and 4 in [10], corresponding to ip vs. IP prediction in the present study. The features used in [10] are full syntactic parse.

Experiment	Ingulfsen (2004)	Current Experiment
Precision of ip	42.9%	45.31%
Recall of ip	5.6%	27.88%
Precision of IP	74.9%	80.00%
Recall of IP	77.9%	86.77%

Table 7: Experimental Results of [9] on the Prediction of Presence or Absence of Prosodic Phrase.

Machine Learner	Accuracy
C4.5	88.8%
SLIPPER	89.8%
QUEST	88.9%
Neural Network	89.2%
Naive Bayes	88.9%

Table 8: The comparison of observed boundary tones with predicted boundary tones. The bold face indicates deviation of the predicted prosodic features from the observed prosodic features. The last column shows the predicted boundary tones with the Festival Speech Synthesis System [8].

Word	Observed	Predicted	Festival
That			
year	IP	**ip**	
Thomas			
Maffy	IP	IP	ip
now			
president	ip	ip	
of			
the			
Massachusetts			
bar			
association	IP	IP	ip
was			
Hennessy's			
law			
clerk	IP	IP	IP

prosodic phrase boundaries. This study shows that the inclusion of linguistic information in modeling prosodic events achieves better accuracy. The approach to prosody prediction developed here holds promise for improving Text-To-Speech (TTS) through more natural prosody, which should enhance intelligibility and naturalness, and has applications as well for prosody detection in ASR [20].

8. Acknowledgment

I am grateful to Jennifer Cole, Mark Hasegawa-Johnson, Richard Sproat, Chilin Shih, Chin-W. Kim, Margaret Fleck, Stefanie Shattuck-Hufnagel, Mark Swerts, Nanette Veilleux, Julia Hirschberg, Pauline Welby, and Lisa Pierce for their useful input and discussions. This work is supported by NSF award number IIS-0414117 and by summer fellowships (2004 & 2005) from the CS/AI at the Beckman Institute, UIUC. Statements in this paper reflect the opinions and conclusions of the author, and are not endorsed by the NSF or University of Illinois.

9. References

[1] Nespor, M.; Vogel, I., 1986. *Prosodic Phonology*, Dordrecht: Foris.

[2] Beckman, M.; Pierrehumbert, J., 1986. Intonational structure in Japanese and English, *Phonology Yearbook* 3, 255-30.

[3] Ladd, D.R., 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[4] Selkirk, E., 2000. The interaction of constraints on prosodic phrasing. In *Prosody: Theory and Experiment*, M. Horne (ed.). Dordrecht: Kluwer, 231-261.

[5] Ladd, D.R., 1986. Intonational phrasing: the case for recursive prosodic structure, *Phonology Yearbook 3*, 311-340.

[6] Steedman, M., 2000. Information Structure and the syntax-phonology interface, *Linguistic Inquiry* 31, 649-689.

[7] Ross, K.; Ostendorf, M., 1996. Prediction of abstract prosodic labels for speech synthesis, *Computer Speech and Language* 10, 155-185.

[8] Taylor, P.; Black, A., 1998. Assigning phrase breaks for part-of-speech sequence, *Computer Speech and Language* 12, 99-117.

[9] Cutler, A.; Dahan, D.; van Doneselaar, W., 1997. Prosody in the comprehension of spoken language: a literature review, *Language and Speech* 40, 141-201.

[10] Ingulfsen, T., 2004. Influence of syntax on prosody boundary prediction, *Technical Report* 610, Computer Laboratory, Cambridge University.

[11] Cohen, A., 2004. *A Survey of Machine Learning Methods for Predicting Prosody in Radio Speech*, Ms.C., Department of Electrical Engineering and Computation, Univeristy of Illinois at Urbana-Champaign.

[12] Ostendorf, M.; Price, P.; Shattuck-Hufnagel, S., 1995. *The Boston University Radio News Corpus*, from <http://www.ldc.upenn.edu>.

[13] Beckman, M.; Ayers, G., 1997. *Guidelines for ToBI Labelling* (version 3.0). ms., The Ohio State University.

[14] Pierrehumbert, J., 1980. *The Phonetics and Phonology of English Intonation*. Ph.D. Dissertation, MIT, Cambridge, MA.

[15] Pierrehumbert, J.; Beckman, M., 1988. *Japanese Tone Structure*. Cambridge, MA: MIT Press. 2729-2732, 2004.

[16] Pierrehumbert, J.; Hirschberg, J., 1990. The meaning of intonational contours in discourse. In *Intentions in communication*, Cohen, P.;Morgan, J.;Pollack, M. (Eds.). Cambridge, MA: MIT Press, 271-311.

[17] Bachenko, J.; Fitzpatrick, E., 1990. A computational grammar of discourse-neutral prosodic phrasing in English, *Computational Linguistics* 16, 155-170.

[18] Welby, P., 2003. Effects of pitch accent position, type, and status of focus projection, *Language and Speech* 46, 53-81.

[19] Daelemans, W.; Zavrel, J.; van der Sloot, K.; Antal van den Bosch, A., 2003. TiMBL: Tilburg Memory Based Learner, version 5.0, Reference Guide. *ILK Technical Report* 03-10.

[20] Chen, K. 2004. *Prosody dependent speech recognition on American radio news speech*. Ph.D. dissertation, University of Illinois at Urbana-Champaign.