Tree Grammars as Models of Prosodic Structure

Joseph Tepperman and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory
University of Southern California, Los Angeles, USA
http://sail.usc.edu/
tepperma@usc.edu, shri@sipi.usc.edu

Abstract

The common ToBI system of transcription assumes a sequential model of prosody. Many linguists argue for a tree structure explaining the synchronization and interaction among prosodic units. Could tree grammars, used previously in syntax-based language modeling, be used to model prosodic trees? We present a method of converting sequential transcripts into trees, and then demonstrate that modeling trees rather than sequences of prosodic tags results in lower perplexity as well as lower error rates when classifying pitch accents and boundaries on the Boston University Radio News Corpus. This finding could benefit areas like speech synthesis, speech understanding, and pronunciation evaluation.

Index Terms: prosody, tree grammar, intonation

1. Introduction

Intonation - the pattern of pitch in speech - progresses sequentially, in a way. If we take intonation to be made of categorical prosodic events (the pitch accents and boundary tones that convey linguistic information through suprasegmental cues such as pitch), then we can imagine one pitch accent following another until an intermediate or final boundary tone intervenes, marking a division in the phrase. This is the model on which a prosodic transcription system like ToBI (Tones and Break Indices) is based [11] - accent and boundary phenomena are modeled not so much as “beads on a string” (as with phonemes), but more like notches on a stick; they are transcribed as discrete events occurring in sequence, denoted by the instants in time perceived as their centers.

Is this really the best way to look at intonation? Consider two phrases, one ending in a low boundary and the other ending in a high one. Anticipation of the boundary motion in each could result in contrasting phrase-level frequency contours, affecting the shape of any within-phrase pitch accents. This is the basis for a superpositional model like Fujisaki’s [3], in which the frequency contour is decomposed into the summation of tone components on the phrase and accent levels. Other theories focus on this nested hierarchy to intonation [5], employing tree structures to schematize the way multiple levels of information are superimposed, and to explain the coordination and synchronization among different scales of prosodic units.

In recent years, tree grammars for sentence syntax have shown some promise in structure-based language modeling for text translation and speech recognition [1, 2]. Tree grammars are capable of capturing long-term context that n-gram models would miss, and are versatile in their modeling of an entire tree as a context-dependent set of subtree structures. Even so, their use has been limited - the ordinary left-to-right decoding of most speech recognition frameworks has favored simpler n-gram language models (especially for real-time processing), and training tree grammars requires expert part-of-speech annotation of sentences. However, tree grammars of prosody rather than syntax do seem well-suited for modeling the sort of structure hypothesized in linguistic theory. The symbol set is small compared to words or part-of-speech tags (potentially requiring less training data), and prosodic tree structures can be derived directly from more common sequential ToBI transcripts.

This study intends to answer two essential questions: given a set of prosodic tags over an utterance, how do we estimate their probability, accounting for all interactions and dependencies? And, secondly, can a tree-based model provide “better” probability estimates than a sequential model?

Perhaps a better question to start with is, why would we want to do this at all? With an estimate of the probability of a set of prosodic tags, we can potentially choose the best set of tags by searching over all possible sets. In speech synthesis this means, for a string of words, we can choose the best prosodic structure to match those words, so that synthesis can sound more natural, with the best pitch accents and boundaries in appropriate places [4]. Using these grammars to decode prosodic tags from acoustic-prosodic features can help to resolve ambiguities in decoding words, or to tag dialog acts for improved speech understanding [10]. They should also be useful in estimating a prosodic pronunciation score for nonnative speakers practicing English as a foreign language - once decoded, a set of prosodic tags common in native speech would receive a higher score than those not common in native speech [12]. Any task that uses categorical prosodic tags (like those in the ToBI system) is a potential application for these tree grammars.

This paper starts off by giving some background on linguistic theories of tree structure in prosody. Then details about data preparation are explained. Next is a short review of tree grammars in general, as they relate to this work. The following section describes some experiments comparing tree-based models with sequential n-gram models. Finally, we draw some conclusions about the benefits of using tree structures for this task.

2. Trees in Prosody

For many years linguists have organized the syllables of English into a hierarchy of prosodic units on various levels, each corresponding to a unique time-scale and function as information [5]. Many claim that the fundamental rhythmic/melodic unit of English prosody is the foot (a term borrowed from studies of poetry), consisting of a stressed syllable and all subsequent unstressed syllables before the next stress. Stress is a syllabic prominence realized through increased duration, energy,

[1] This work was supported in part by the NSF and the NIH.

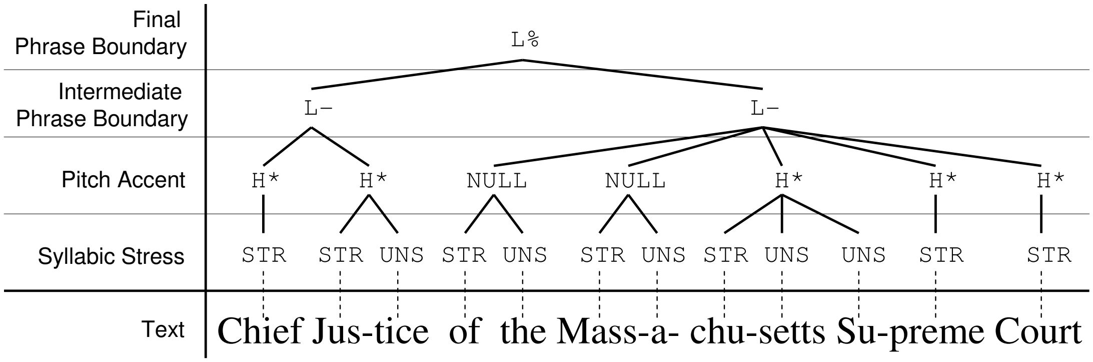

Figure 1: A prosodic tree derived from sequential ToBI transcripts and syllable segmentation of the BURNC [8]. The four tiers of the tree represent prosodic units on increasingly large time-scales, each with a unique linguistic meaning. The ASCII representation of the top three tiers would be `TOP(L%(L-(H* H*) L-(NULL NULL H* H* H*)))`. The text is not part of the tree structure but is included here for illustration. Note that the pitch accents' bounds match the foot rather than the word, following [5].

	train set	*dev set*	*test set*
total tags	37792	4973	4685
total trees	3756	493	465
total minutes	62.0	8.1	7.7

Table 1: Sizes of the training, development, and test sets. "Tags" refers to the prosodic symbols in the transcripts. "Trees" means complete four-level prosodic trees.

or pitch, and is used to mark lexical contrasts between words (e.g. "contract"). On the level of the prosodic foot, pitch accents - manifested through changes in pitch and energy - are perceptually relevant to discourse, denoting if the information offered is new, contrastive, accessible, or uncertain [14]. Above that, phrase boundary tones on the intermediate and final levels act like prosodic punctuation, offering the listener cues for interpretation and syntactic processing [13]. All of these levels of information are present simultaneously in the natural pitch and energy, and their organization is schematized well as a tree structure to illustrate the way the units nest and co-occur.

Uses of these hierarchical theories in computational models of prosody have been relatively rare. A few studies in syntax-based language modeling have combined syntactic and prosodic trees, with improvements in predicting boundary locations [4, 7]. One study in French prosody used a tree structure of an entirely different kind, based on syllable grouping according to pitch range and slope [9]. Generally, most modeling of intonation has remained sequential rather than tree-structured, in line with the ToBI system [11] that seems to dominate prosodic transcription.

3. Corpus Preparation

All prosodic tag data in this study come from the ToBI annotation of the Boston University Radio News Corpus (BURNC) [8], which consists of read news reports by professional radio announcers - the intonation they employ is highly regularized and is representative of a generic standard for American read speech. In addition to ToBI labels, the BURNC has transcripts for syllable-level stress as well as syllable boundary locations. All transcripts from one speaker were divided into training, development, and test sets - their sizes can be found in Table 1.

To define tree structures, we needed to align all levels of annotation with shared beginning and ending boundaries, though in the transcripts only the syllable boundaries were defined. The end of an intermediate phrase we defined as the end of the syllable in which the boundary tone's center was transcribed; its beginning was either the beginning of the utterance or the end of the previous intermediate phrase. Full phrase boundaries were determined the same way, and since full boundaries require a concurrent intermediate boundary, this syllable demarcation synchronized the two boundary levels and established the tree structure between them. Following [5], we assume a pitch accent in English extends throughout its prosodic foot. We defined the foot as beginning with the syllable in which the accent's center was transcribed, which was ordinarily the foot's stressed syllable (but occasionally fell on an unstressed syllable in the BURNC). The end of that foot was then defined as synchronous with the beginning of the next stressed syllable, intermediate phrase, or silence. All leftover unaccented feet within a phrase were assigned a `NULL` accent tag. This established the nesting of stressed or unstressed syllables within a pitch-accented prosodic foot, and of the pitch accents within the intermediate phrases. See Fig 1 for an example of a derived tree. The top three tiers, when put in sequential order, would be {`H* H* L- NULL NULL H* H* H* L- L%`}.

The full and intermediate phrase boundaries can each take high and low symbols - {`H%,L%`} and {`H-,L-`}, respectively - and because of low inter-annotator agreement for fine-grained pitch accents, these were distilled down to two categories, `H*` and `L*`, in addition to the accentless `NULL`. Similarly, all syllable stress labels were binary, `STR` or `UNS`, with secondary stress within a word considered simply `STR`. Any transcribed silence between phrases was assigned the tree `TOP(SIL1(SIL2(SIL3(SIL4))))` in keeping with the four tiers of sequential symbols within the tree. Rarely, some of the accent or boundary tone labels were unspecified in the transcripts due to uncertainty on the transcriber's part - these we mapped to the most common tags, `H*` and `L-`, respectively.

4. Regular Tree Grammars

A Probabilistic Context Free Grammar (PCFG) specifies a set of terminal and nonterminal symbols for which, beginning with a starting symbol, a sequence of tree production rules for re-

	n-gram order					*trees*	
	1	*2*	*3*	*4*	*5*	*PI*	*PD*
dev	2.72	2.53	2.49	**2.48**	2.49	2.29	**2.27**
test	-	-	-	**2.41**	-	-	**2.27**

Table 2: PPL results for 4-tier setup, over dev and test sets.

	n-gram order					*trees*	
	1	*2*	*3*	*4*	*5*	*PI*	*PD*
dev	3.80	2.74	2.63	**2.61**	2.63	2.46	**2.44**
test	-	-	-	**2.53**	-	-	**2.44**

Table 3: PPL results for 2-tier setup, over dev and test sets.

placing the nonterminal symbols can be performed, each with an associated probability [6]. The probability of a tree T is then the product of the probabilities for all n production rules $\alpha \rightarrow \beta$ that generated it,

$$P(T) = \prod_{i=1}^{n} P(\alpha_i \rightarrow \beta_i | \alpha_i) \qquad (1)$$

This of course assumes all rules (and all subtrees generated by those rules) are independent, allowing for the versatility of modeling a larger tree implicitly through a sequence of smaller tree production rules. A Weighted Regular Tree Grammar (WRTG) is a finite-state acceptor of PCFG trees, representing all nonterminal symbols through states in the recognition network. Probabilities of production rules (i.e. state transitions through the WRTG) are estimated from the training data as

$$\hat{P}(\alpha \rightarrow \beta | \alpha) = \frac{Count(\alpha \rightarrow \beta)}{Count(\alpha)} \qquad (2)$$

where $Count(\alpha \rightarrow \beta)$ and $Count(\alpha)$ are the occurrences of the production rule $\alpha \rightarrow \beta$ and the symbol α, respectively.

In the case of the prosodic trees being modeled here, it's clear that our starting symbol is TOP, and our only terminal symbols are {STR,UNS,SIL4}. Unlike the parse trees for part-of-speech tags, our prosodic trees are simpler in that our nonterminal symbols are not recursive. No nonterminal prosodic tag can produce itself in the tree the way that certain parse tags (like the noun phrase, NP) can - for example, there can never be an H* inside of an H*, but there can be noun phrases inside of noun phrases.

Equation 1 defines how $P(T)$ is calculated when the tags in T are arranged in a tree structure. When taken sequentially, a traditional n-gram model estimates $P(T)$ as

$$P(T) = P(t_1, \ldots, t_{|T|}) = \prod_{m=1}^{M} P(t_m) \qquad (3)$$

where t_m is one of M parallel but independent sequences of symbols in the set T (e.g. tiers of the tree). A separate n-gram model for each sequence defines $P(t_m)$ as

$$P(t_m) = P(s_m^1, \ldots, s_m^Q) = \prod_{q=n}^{Q} P(s_m^q | s_m^{q-1}, \ldots, s_m^{q-n+1}) \qquad (4)$$

where q is the symbol index in the sequence, and n is the order of the n-gram model.

5. Training and Experiments

The main thing we wanted to learn from this study was whether we can improve models for prosodic tags using tree grammars instead of sequential n-gram models. "Improvement" we will measure using perplexity (PPL), the standard metric for comparing two different language models' abilities to assign high probabilities to previously unseen strings of symbols or words. Ultimately, classification or detection error rates offer a truer comparison, but PPL is a useful metric to estimate relative model performance. In our case, the PPL is defined as

$$PPL = 2^{-log_2\{P(T)\}/|T|} \qquad (5)$$

where $P(T)$ is the probability the model assigns to the set of tags T, and $|T|$ is the number of tags in the set.

For the sequential n-gram models, the performance was evaluated over two different experimental setups. One, the "4-tier" setup, took the four tiers of prosodic tags to be independent parallel sequences. The other, the "2-tier" setup, combined the top three tiers into a sequence encompassing pitch accents, intermediate boundary tones, and final boundary tones, just like in the ToBI transcripts. The syllable stress tier had to remain separate since many syllables are simultaneous with the pitch accents and boundary tones, rather than sequential. The tree grammars were identical in each case, but to keep the number of tags the same, the silence trees for the two-tier setup were changed to TOP(SIL1(SIL4)). It should go without saying that the TOP symbols were not included in the tree perplexity calculations, since they are not in the sequential transcripts.

Similarly, two types of tree grammars were trained. Parent-dependent (or PD) tree grammars were conditioned on knowledge of the parent symbol one level up in the tree. For example, instances of the production rule {H* → STR UNS} would be split into either {H*|L- → STR|H* UNS|H*} or {H*|H- → STR|H* UNS|H*}, depending on H*'s parent. Parent-independent (or PI) tree grammars simply did not use this knowledge of the parents, intead assuming, for example, that all examples of the production rule {H* → STR UNS} were to be modeled as one, regardless of H*'s parent.

The best sequential model for each setup was found by increasing the order of the n-gram until perplexity on the development set no longer decreased. These n-grams were trained using the SRI language modeling toolkit with Good-Turing smoothing. All tree grammars were implemented using Tiburon [6], with probabilities trained using the method in Eqn. 2. The subtree production rules in the dev and test sets not seen in the training set were assigned a count of 1 before all probabilities were normalized - this is known to be a simple and sub-optimal smoothing method. The better of the two tree grammars on the dev set was evaluated on the test set, for comparison with the best n-gram model. These results are given in Tables 2 and 3.

As a preliminary experiment beyond perplexity, we also did prosodic tag classification by answering this question: given all tags but one in a tree or sequence, what should that missing tag be? The classification result for a set of tags T was given by

$$\hat{R} = \underset{R}{\operatorname{argmax}} \{P(t_1, t_2, \ldots, R, \ldots, t_{|T|})\} \qquad (6)$$

where R can be any other tag from the missing tag's tier. The intermediate and final phrase boundary tone classification was binary - either low or high - and classification of pitch accents

	# test items	*majority choice*	*4-tier 4-grams*	*2-tier 4-grams*	*PD trees*
FINB	*275*	28.33	35.84	29.69	**24.92**
INTB	*389*	21.88	20.19	15.63	**11.78**
PACC	*1203*	46.22	33.59	27.04	**24.58**

Table 4: Tag classification error (in %) on the development set. *FINB* = final boundary tones, *INTB* = intermediate boundary tones, *PACC* = pitch accents.

was three-way: H*, L*, or NULL. Essentially this is very much like the speech synthesis task of assigning natural-sounding prosody to text - we decide what type of accent or boundary tone to have, given that we already know where the accent or boundary should be (perhaps based on the syntactic phrases, or the lexically-defined syllable stress sequence). Results for this classification are given in Table 4, using the best n-gram and tree grammars from the PPL experiments, alongside a "majority choice" baseline in which all test items were assigned the most common tag.

6. Discussion

In the PPL experiments (Tables 2 and 3) we see that the best models for both setups were 4th-order n-grams and PD trees, with the PD and PI trees outperforming the best n-gram models. This indicates that the long-term and multi-scaled context captured by the tree grammars is better suited for modeling prosodic structure, and is evidence in favor of the linguistic theories on which the tree structures are based. The improvement seen with context-dependent models in both the sequential and tree grammars illustrates the importance of accounting for as much of the prosodic structure as possible. However, the difference in PPL between the PI and PD trees was not as large as between the 2nd and 3rd order n-grams, suggesting that vertical dependencies in the tree are not as important as horizontal ones. With more training data, higher-order n-grams might yield some improvement, but more training data would probably make the tree grammars better as well. In general the perplexity for all models was quite low, due to the small set of prosody labels.

The 2-tier and 4-tier setups are not really directly comparable in the PPL experiments because they have different symbol sets (due to combination of SIL tags). For the tag classification experiments reported in Table 4, we see that the 2-tier sequential setup outperformed the 4-tier one for all three tag types, indicating that the classification of pitch accents can benefit from knowledge of phrase boundaries, and vice versa - one assumption that motivated our use of tree grammars to begin with. As for the PD tree grammars, they outperformed the sequential grammars and "majority choice" baseline in all cases, with a margin of 2-5% over the best n-grams. Improvement over the baseline with PD trees was most dramatic for the pitch accents, partly because the three-way classification made for higher baseline error. Neither of the sequential methods beat the baseline for the phrase-final boundaries, suggesting independence between successive phrase-final tones in the sequence.

7. Conclusion

These experiments have shown that using tree grammars to model the structure of prosodic tags has several advantages over sequential models, including lower perplexity measures and lower prosodic tag classification error rates. This seems to justify the linguistic theories behind schematizing prosody in a tree structure, and is potentially useful in applications as diverse as speech synthesis, dialog act classification, and pronunciation scoring. The next step would be to combine these tree-based "language models" for a set of prosodic tags with acoustic models for the suprasegmental manifestations of those tags, so that they can be decoded from speech. To do this within the framework of traditional left-right speech decoding will be challenging, and is one potential drawback of tree grammars.

8. References

[1] E. Charniak, K. Knight, K. Yamada, "Syntax-based Language Models for Statistical Machine Translation," in *MT Summit IX*, Int'l. Assoc. for Machine Translation, 2003.

[2] C. Chelba and P. Xu, "Richer Syntactic Dependencies for Structured Language Modeling," in *Proc. of ASRU*, 2001.

[3] H. Fujisaki, "Prosody, Models, and Spontaneous Speech," in *Computing Prosody*. Ed. Y. Sagisaka, N. Campbell, and N. Higuchi. New York: Springer, 1997.

[4] J. Hirschberg, O. Rambow, "Learning Prosodic Features Using a Tree Representation," in *Proc. of Eurospeech*, Aalborg, 2001.

[5] D. Hirst, "Intonation in British English," in *Intonation Systems: A Survey of Twenty Languages*. Ed. D. Hirst and A. Di Cristo. Cambridge: CUP, 1998.

[6] J. May and K. Knight, "Tiburon: A Weighted Tree Automata Toolkit," in *Proc. of the Eleventh Conference on Implementation and Application of Automata*, 2006.

[7] S. Minnis, "The Parsody System: Automatic Prediction of Prosodic Boundaries for Text-to-speech," in *Proc. of the International Conference on Computational Linguistics*, Kyoto, Japan, 1994.

[8] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The Boston University Radio News Corpus," Boston University Technical Report No. ECS-95-001, March 1995.

[9] N. Segal and K. Bartkova, "Prosodic Structure Representation for Boundary Detection in Spontaneous French," in *Proc. of ICPhS XVI*, Saarbrucken, 2007.

[10] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech*, 41(3-4), 439-487, 1998.

[11] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A Standard for Labeling English Prosody," in *Proc. of ICSLP*, Banff, Canada, 1992.

[12] J. Tepperman, A. Kazemzadeh, and S. Narayanan, "A text-free approach to assessing nonnative intonation," in *Proc. of InterSpeech ICSLP*, Antwerp, Belgium, August 2007.

[13] J. Vaissiere, "Perception of Intonation," in *The Handbook of Speech Perception*. Ed. D. B. Pisoni and R. E. Remez. Oxford: Blackwell, 2005.

[14] A. Wennerstrom, *The Music of Everyday Speech*. New York: OUP, 2001.