MARKOV MODELING OF PROSODIC PHRASE STRUCTURE

N. M. Veilleux — Boston University, Boston, MA 02215
M. Ostendorf — Boston University, Boston, MA 02215
P. J. Price — SRI International, Menlo Park, CA 94025
S. Shattuck-Hufnagel — Mass. Inst. of Technology, Cambridge, MA 02139

Abstract

Good models of prosodic phrase structure are needed for both higher quality text-to-speech synthesis and for parsing in speech understanding. In this paper, we describe a simple computational model for predicting phrase boundaries from text. The hierarchical structure of the model is based on linguistic theory, and the model itself is probabilistic. Results are presented for the model, trained and evaluated on a database of FM radio news.

I. Introduction

Prosodic phrases are sequences of words in speech that are perceived as being grouped. Such groupings are often signaled by intonation (F0) and duration cues. A computational model of prosodic phrase structure is necessary for high quality text-to-speech synthesis, where correct assignment of phrase breaks can increase the intelligibility of a sentence as well as improve its naturalness. A prosodic phrase model may also prove useful for speech understanding systems; although prosodic structure and syntactic structures are not completely identical, prosodic phrase information may be useful in resolving some syntactic ambiguities [1].

This paper considers computational modeling of prosodic phrase structure given text, i.e., a prosodic parser. Past work in prosodic parsing has mainly focused on speech synthesis and initially involved syntactic parsing [2]. More recent approaches have been based on the hypotheses that 1) a prosodic parse may not require a full syntactic parse and 2) detailed part-of-speech information (e.g., noun, verb, determiner) may not be necessary for generating a prosodic parse. Sorin *et al.* [3] proposed a prosodic parser for French based on content/function word classification and a few simple rules. An advantage of this approach is that it only requires a small dictionary of function words to assign part-of-speech labels. O'Shaughnessy [4] proposes a somewhat more sophisticated parser for English, motivated by similar principles.

This work was supported by the National Science Foundation, grant number IRI-8805680.

Our approach is based on the same principles and is an extension of the French work. The model has a linguistically motivated, discrete component and a probabilistic component. The probabilistic component is important for several reasons: it can capture the fact that the same sequence of words can be produced with a variety of prosodic contours (some more likely than others), it allows for automatic training of the model, and it provides a mechanism for combining a prosodic component with other knowledge sources.

In the remainder of the paper, the prosodic phrase model is described in more detail, initial experimental results are presented, and possible extensions of this work are discussed.

II. Computational Model

Recent proposals in linguistic theory share the view that prosodic structure includes at least a 2-level hierarchy of major and minor phrases. For example, Beckman and Pierrehumbert [5] describe intonational phrases made up of smaller intermediate phrases, cued acoustically by sequences of pitch accents, as well as pauses and duration lengthening at intonational phrase boundaries. A more complex hierarchy that groups intonational phrases has been suggested by Ladd [6]. A detailed hierarchy of smaller prosodic units can also be described; but our current goal is to predict major and minor phrase constituents, with the initial effort focusing on minor (intermediate) phrases.

The model described here is a two-level hierarchy of minor phrases comprised of "prosodic groups", which are similar to the units described in [3]. A third level can be added to model major phrases, or intonational phrases, as described later. Unlike the minor and major phrases, the prosodic groups are not associated with a phonological theory, but are used to simplify prediction of the higher level groups.

Model Description and Parameter Estimation

The prosodic phrase model assigns probabilities to different prosodic parses for a given input sentence. The components of the model are illustrated in Figure 1: part-of-speech labeling, prosodic group assignment, Markov phrase modeling.

CH2847-2/90/0000-0777 $1.00

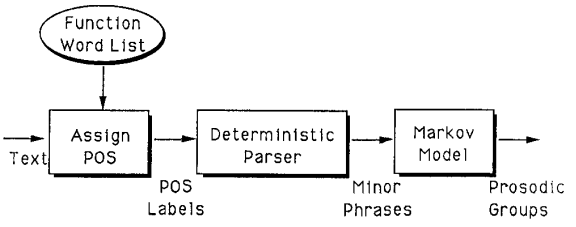

Figure 1: Block diagram of prosodic phrase model.

Part-of-Speech Assignment. Given the text of a sentence, the first step involves converting the sequence of words to a sequence of content/function part-of-speech (POS) labels. There are five POS classes: auxiliary verb (v), subjunctive and conjunctive function words (s), all other function words (f), proper nouns (C), and all other content words (c). Words are assigned POS labels by searching the three function word lists and testing for capitalization. All words which are not assigned using these rules are classified as content words (c).

Prosodic Groups. Given the POS sequence, prosodic groups are assigned using a slight modification of the rules for French [3]. Prosodic group boundaries are placed after each content word that is followed by a function word, at the beginning and end of a sequence of proper nouns, and (optionally) at sites indicated by commas or other punctuation marks. After the prosodic group boundaries are marked, the sequence of words comprising each group is assigned a label according to the POS assignment of the first word in the group. Although this algorithm was successful for defining prosodic groups in French, it seemed to generate too many boundaries for English. However, the algorithm was useful for marking candidate locations for phrase boundaries in English and could be thought of as a "data compression" step in our model.

Minor Phrases. Minor phrase boundaries are located by modeling the sequence of prosodic group labels and phrase breaks as a six-state Markov chain. The six states are the five prosodic group labels (v, s, f, C, c) and a phrase break label (b). All possible transitions between states are allowed, with the exception of transitions from a break to another break. Both first- and second-order Markov models were investigated. The low order Markov models actually represented relationships of fairly long word sequences because of the prosodic group compression step.

The parameters of the model are the probabilities of transitions between states: $p(s_n|s_{n-1}, s_{n-2})$ or $p(s_n|s_{n-1})$. Maximum likelihood estimates of these parameters are relative frequencies of three-state sequences (trigrams) and two-state sequences (bigrams) appearing in the training data. While the first-order transition states were covered by the training data, the representation of second-order transitions was sparse. Therefore the transition probabilities were calculated as a mixture of first and second order relative frequencies.

The Markov minor phrase model is similar to Markov language models used in speech recognition, except that the minor phrase model represents sequences of prosodic group labels and phrase break markers rather than sequences of words. In addition, 'b' states are unobserved when the model is used to predict phrase boundaries, as described in the next section.

Phrase Modeling in Synthesis and Analysis

The model structure computes a probability for any parse allowable from the prosodic group candidate breaks, but these probabilities can be used in several different ways depending on the application (synthesis or analysis). The application also affects the assumptions that can be made about the input text. In most text-to-speech synthesis applications, punctuation is known and can be incorporated into the model by: 1) using a rule that includes a candidate break at every comma and 2) deterministically assigning minor phrase breaks at commas. In analysis applications such as speech understanding, however, punctuation is not known and these deterministic rules are not applied.

In text-to-speech synthesis, a prosodic phrase model would be used to generate a single prosodic parse for each sentence. The prosodic parse is then used to determine pitch range and placement of boundary tones and duration lengthening. Using the model described above, minor phrase boundaries are located by first assigning POS of each word and then determining the prosodic groups. The sequence of prosodic

```
When a computerized call  is made
  s  f      c        c    v   c
(           s          )(   v    )
[<                                >

to a former prisoner's home phone
f  f  c        c       c     c
(               f               )
<                               >]

that person answers  by plugging
 f     c       c     f     c
(         f        )(     f     )
[<                 ><

in the device.
f   f    c
(     f       )
             >]
```

Figure 2: Example of the various stages of a prosodic parse. The first line is the text input; the second line shows the POS assignment; the third line gives the prosodic group classifications; and the fourth line shows the minor phrase boundaries < > and hypothetical major phrase boundaries [ ].

group labels, together with phrase boundary labels known because of punctuation, then represents a partially observed Markov chain. The unknown minor phrase breaks are the unobserved states in the sequence. These unknown breaks are predicted either deterministically, by finding the most likely locations, or randomly, by choosing state transitions according the the Markov model parameters. An example of the different stages of output for a prosodic parser is given in Figure 2.

Speech understanding is an analysis application which might benefit from the prosodic phrase model. One possible scenario would involve assigning probabilities to candidate prosodic parses, and multiplying the probability of a parse given the word sequence with the probability of the parse given the acoustic evidence. The combination of the two knowledge sources should provide a more reliable score of the prosodic phrase boundaries, in the same manner as Markov language models are used to improve speech recognition accuracy. The prosodic parse information would then be supplied to a syntactic parser to eliminate any prosodically inconsistent hypotheses.

III. Experimental Results

Database

The database for training and evaluation is recorded FM radio news broadcasts by professional speakers. Minor and major phrase boundaries in the speech were hand-marked by at least two listeners. Three stories from two different announcers (92 sentences total) were used for training the model, and a fourth story (22 sentences) was used to evaluate the model. The evaluation story was read five times, by four different announcers, in order to capture some of the possible prosodic variation. Listeners heard an average of 2.75 parses/sentence among these five versions, where a parse is considered different if either minor or major phrase boundaries differed. Certainly, there are other acceptable ways to say these sentences, but the five versions provide some indication of the variety associated with the radio style.

Deterministic parser

The assignment of prosodic groups by the deterministic parser affects all higher levels of phrase boundaries, because the prosodic groups define the *candidate sites* for phrase breaks. Consequently, this algorithm must have a high rate of correct prediction of breaks, at the cost of some overgeneration to allow different prosodic parses. The rules described here define small prosodic groups, with an average size of 2.6 words in the test story. The performance of the deterministic parser is evaluated by comparing the perceived phrase breaks in the spoken sentences of the five test stories to the location of the candidate boundaries predicted by the model. An average of 89% of the spoken boundaries were predicted by the deterministic parser for the five versions of the story when punctuation rules (commas) were used (84% without punctuation rules), and 43% of the candidate boundaries were not used in any of the five versions of the test story. Half of the boundaries that were not detected were used in only one of the five versions, and many (8/21) were due to boundaries incorrectly predicted before a particle rather than after. (See "plugging in" in Figure 2, as an example.) The particle-preposition ambiguity of many words in English is an area where we expected the algorithm to have difficulty. This problem will need to be addressed in future work, possibly by doing trigram modeling of word sequences based on hand labelings to do better prediction of the particle or preposition POS for many function words. There was only one location where all five announcers put a boundary and the deterministic parser did not. The error could have been corrected had a comma appeared there in the original text: "... fifty times a week according to Ash, who says ..." vs. "... fifty times a week, according to Ash, who says ..."

Minor Phrase Breaks

Since the way the Markov model is used depends on whether the application is synthesis or analysis, different techniques for evaluating the minor phrase model are presented below.

For synthesis applications, the minor phrase parse was given by the most probable sequence of prosodic groups and phrase breaks, assuming knowledge of punctuation. Each automatically parsed sentence was compared to the same sentence in the five versions of the test story. Two of the sentences could not match any of the test versions because of missing break candidates. Of the 20 remaining sentences, 9 were automatically assigned a parse that was spoken in one of the versions. While this is not a decisive indication that the model output would sound natural, it is expected to be a lower bound on estimated naturalness. Inspection of the produced parses led us to believe that five other sentences had acceptable parses. Using the second order Markov model did not improve the results, but the use of commas was important (without commas: 7/20 sentences correct, 3 acceptable). Parsing was particularly difficult for this text because many of the sentences were long, and more careful use of punctuation in the transcription would have improved performance further.

In analysis applications, the function of the prosodic phrase model is to weight scores or restrict a recognition search space according to the probability of different prosodic parses. These probabilities can also be used to aid in evaluation of the model. Ideally, parses used by the announcers should be rated highly probable. In the 22 test sentences, an average of 11 per speaker were spoken with parses that occurred among the top 10 most probable sequences according to the model. (Under this measure, the second-order model produced slightly better results.) 60% of the sentences not

in the top ten were eliminated because of missing boundary candidates. Most other sentences that were not in the ten most probable were either very long (over 8000 possible parses) or included parenthetical phrases. Another measure of a language model is perplexity, which is effectively the average branching factor from a state. With pause constraints and equally probable transitions, the branching factor of the six-state model is 5. The test set perplexity of the first and second order Markov models is 3.3 and 3.1, respectively.

IV. Conclusions

In summary, we have developed a hierarchical model for predicting and scoring prosodic phrase structure, representing intonational structures that are also reflected in current linguistic theory. This model has several advantages: it is simple to implement (relative to a syntactic parse), represents prosodic variation, and can be used in both analysis and synthesis. The current version models minor phrase boundaries, but can be easily extended to include major phrases and other higher level units. One approach involves adding another state to the Markov chain, but our initial results with this method were only marginally successful (6/20 correctly predicted parses). Alternatively, a separate model could represent the probabilistic relationship between sequences of minor phrases and major phrase breaks as in [7].

The experimental results with this model were encouraging, but it is clear that much more data must be used for training and testing. It might also be interesting to consider speaker-dependent models. In addition, the types of errors observed in generating parses for synthesis suggest that use of commas is very important for parsing long sentences, and in analysis, the spoken sentences which did not have probable parses suggest that it might be useful to separately model parenthetical breaks. Both tests indicated that the deterministic parser needs to be improved, especially in handling particles. Finally, this model used very coarse POS assignments. It would be interesting to investigate the improvements in the algorithm associated with a more detailed POS classification method in conjunction with a respecification of the deterministic parser rules.

References

[1] P. J. Price, M. Ostendorf and C. W. Wightman, "Prosody and Parsing," in the *Proceedings of the Second DARPA Speech and Natural Language Workshop,* October 1989.

[2] J. Allen, S. Hunnicutt, R. Carlson, B. Granstrom, "MITalk-79: The 1979 MIT Text-to-Speech System," *Speech Communications Papers Presented at the 97th Meeting of the ASA,* ed. J. Wolf and D. Klatt, pp. 507-510 (1979).

[3] C. Sorin, D. Larreur and R. Llorca, "A Rhythm-based Prosodic Parser for Text-to-Speech Systems in French," in *Proceedings of the International Congress of Phonetic Sciences,* Vol. 1, pp. 125–128, Tallinn (1987).

[4] D. O'Shaughnessy, "Parsing with a Small Dictionary for Applications such as Text-to-Speech," *Computational Linguistics,* Vol. 15, No. 2, pp. 97-108 (1989).

[5] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook 3,* ed. J. Ohala, pp. 255-309 (1986).

[6] R. Ladd, "Intonational phrasing: the case for recursive prosodic structure," *Phonology Yearbook 3,* ed. J. Ohala, pp. 311–340 (1986).

[7] J. R. Rohlicek, "Statistical Language Modeling Using a Small Corpus from an Application Domain," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing,* pp. 267–270 (1988).