# THE *PARSODY* SYSTEM : AUTOMATIC PREDICTION OF PROSODIC BOUNDARIES FOR TEXT-TO-SPEECH

Stephen Minnis

Natural Language Group, Systems Research Division, BT Laboratories, Martlesham Heath, Ipswich IP5 7RE, UK
tel +44-473-645384, e-mail : sminnis@bt-sys.bt.co.uk@uucl

*As commercial text-to-speech systems move above the word level to the sentence level, the prediction of the correct prosodic information becomes a significant factor in the perceived naturalness of the synthesised speech.*

*This article reports on "**Parsody**", an experimental system which combines a partial parser with a prosodic marking component to predict the location and relative strength of prosodic boundaries in text.*

## INTRODUCTION

Modern text-to-speech (TTS) systems are quite good at word level synthesis, but tend to perform badly on connected word sequences. It has been suggested that the poor prosody of synthetic connected speech is the primary factor leading to difficulties in comprehension [1,5]. TTS systems must therefore incorporate better mechanisms for prosodic processing. For the purpose of this article, prosodic processing is narrowly interpreted as being the prediction of the location and of the relative strengths (salience) of prosodic boundaries (although, of course, there are several other important aspects to prosody). A prosodic boundary is a point in a spoken utterance associated with important acoustic prosodic phenomena, such as pauses and pitch change.

There are two main approaches to the prosodic marking problem : the rule-based approach and the stochastic-based approach.

The rule-based approach stems from Gee and Grosjean's work on performance structures[1] [7], which has been the focus of many extensions, such as that reported by Bachenko and Fitzpatrick [2,4]. Gee and Grosjean's work sought to account for the (then) disparity between linguistic phrase-structure theories and actual performance structures produced by humans, and focused on recreating the pause data of several analysed sentences from syntax (although they claim that their method could easily account for other prosodic features). The central tenet of their work was that prosodic phrasing is a compromise between the need to respect the linguistic structure of the sentence and the performance aspect (which manifests itself as a need to balance the length of the constituents in the output).

More recent efforts have extended the Gee and Grosjean approach in various ways. Bachenko and Fitzpatrick take a similar rule-based approach but believe that syntax plays a lesser role in determining phrasing, and that certain prosodic performance constraints, such as length, override syntactic structure. They allow prosodic boundaries to cross syntactic boundaries (under certain conditions), whereas Gee and Grosjean viewed their rules as acting within basic sentence clauses. Bachenko and Fitzpatrick made several other changes to the basic Gee and Grosjean algorithm, including counting phonological words[2] rather than actual words when determining node strengths. Wightman et al. have proposed some further interesting extensions to the Bachenko and Fitzpatrick method [12].

With the availability of large and accurately labelled prosodically annotated corpora, the stochastic-based approach will come more to the fore. Wang and Hirschberg [11], and Ostendorf et al. [9], have both described methods for automatically predicting prosodic information using decision tree models. Generally, decision trees are derived by associating a probability with each potential boundary site in the text, and relating various features with each boundary site (e.g. utterance and phrase duration, length of utterance (in syllables/words), positions relative to the start or end of the nearest boundary location etc.) [11]. The resulting decision tree provides, in effect, an algorithm for predicting prosodic boundaries on new input texts.

It is interesting to note that Ostendorf et al. report similar results in their evaluations of the performance of both the rule-based and decision tree algorithms.

## THE *PARSODY* SYSTEM

Our approach has been to implement a rule-based method on top of a chart parser. This was a purely practical decision, as an efficient chart parser had been developed in-house. Also the larger and more detailed descriptions of the rule-based methods in the literature provided an adequate starting point on which to build. The resulting system, the *Parsody* (from **Parser** + **Prosody**) system is designed to provide a test-bed for investigating the interface between syntactic parse structures and the performance structures of actual speech. Results from the *Parsody* system are directly fed into BT's TTS system, Laureate.

The fact that Laureate will be a commercial TTS system places several requirements on the parser, it must be robust, it must be fast, and it must predict prosodic boundaries with a reasonable degree of accuracy. At present the emphasis is on the prediction of the location of prosodic boundaries rather than on the strength of the boundaries.

The Parsody system is implemented in C, under X-windows on a Sun Sparc station, and allows for interactive editing of intermediate results throughout the parsing/prosodic marking process. This provides us

[1] Performance structures are "structures based on experimental data, such as pausing and parsing values" [7].

[2] A phonological word is one which effectively functions as one spoken item, as the internal word-word boundaries are resistant to pausing [7]. Typical examples are determiner-noun word groups, such as "the+man".

with a useful tool for investigating our algorithms, as well as a debugging aid.

A description of the main aspects of the parser and the prosodic marking components is now given.

## THE PARSER COMPONENT

It is interesting to note that one of the sentences in the Bachenko and Fitzpatrick appendix of sentences was not parsed because of "too many parse problems". Obviously this is not acceptable for a commercial text-to-speech system. The *Parsody* parser is designed always to produce one result through a combination of stochastic word tagging and partial parsing with a minimal grammar. All processing is performed on a chart data structure back-bone incorporating packing. This overall approach results in a very fast and efficient parser.

A word's part-of-speech is important for TTS as it may affect the word's pronunciation. Stochastic word tagging enables the parser *always* to choose one word tag, although this may or may not be the correct one (the current accuracy is approximately 95% correct - this figure being given on the Bachenko and Fitzpatrick sentences and on other test sentences). Fortunately for pronunciation purposes, the number of words having multiple pronunciations is quite small - between 1 and 2% of words in our lexicon. Initial investigations have shown that there is less than a 0.3% chance of picking the wrong pronunciation for a word.

Another important aspect of the *Parsody* word-tagging approach is that ill-formed input can be accommodated, and the prosodic marking component can still function to produce a result. Some speech is better than none, even if it sounds strange.

The minimal grammar also helps the parser to produce only one, and always one, output. The grammar is a simple LNP/PP grammar augmented by special 'partition' rules. An LNP is simply a 'longest noun phrase' which is an unambiguous interpretation of the longest NP in the parse result. A PP is a prepositional phrase. An example of a partition node is one which is inserted between two immediately adjacent 'longest NPs' in the parse structure[3].

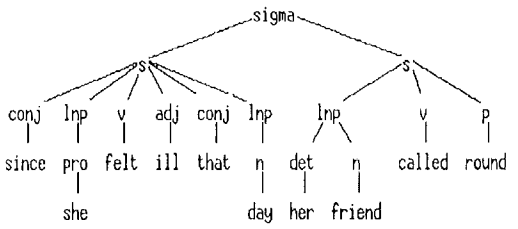


Figure 1. Example Parse Tree produced by *Parsody*

[3]For this reason, each of the prosodic rules to be described works *within* the partition nodes. This is because the boundary between each partition node seems to mark the largest boundaries within the sentence, and the later in the analysis they are joined, the larger the prosodic boundary will be. It may well be that some analysis, perhaps verb adjacency, should take place across partition node boundaries. Further research will examine this.

A typical tree is shown in Figure 1. The partition nodes are denoted by the two "s" labels in this tree.

The partial syntactic tree is then passed to the prosodic marking system.

## THE PROSODY COMPONENT

The prosodic marking algorithms are founded on the Bachenko and Fitzpatrick extensions to the Gee and Grosjean rules.

There are essentially two main components in the Bachenko and Fitzpatrick model. The first, concerning boundary location is basically adhered to in *Parsody*. Boundary location entails the grouping of words into phonological words, and then into phonological phrases. The boundaries separating prosodic phrases form potential prosodic boundary location sites.

The second component seeks to determine the boundary strengths via a series of rules. Bachenko and Fitzpatrick describe a verb-balancing rule which attempts to balance material around a verb, and a verb adjacency rule which in effect extends the verb balancing rule, using 'bundling' (the adjoining of adjacent phrases) to continue to centre material round a verb. Here, *Parsody* employs two main departures from the Bachenko and Fitzpatrick rules. The first is in the domain of verb adjacency. *Parsody*'s verb adjacency algorithm retains the notion of grouping nodes to form a balanced tree, but extends this rule to cover *all* nodes (with the exception of the very final PP). The basic algorithm is also different.

By extending the grouping of nodes to cover all nodes, the confusion of Bachenko and Fitzpatrick's "general bundling rule" is avoided, since all nodes will have been grouped at completion. The change to the algorithm is more subtle, yielding the rule:

**if Count(X) + Count(Y) < Count(Z)**
**then**
**Join to the Left(Y)**
**else**
**Join to the Right(Y)**

**where :**

| | | |
|---|---|---|
| **Count(a)** | = | **Number of Phonological Words beneath Node 'a'** |
| **X** | = | **Previous Node** |
| **Y** | = | **Current Node** |
| **Z** | = | **Next Node** |

This makes explicit the assumption in Bachenko and Fitzpatrick's algorithm that the adjoining of phrases produces a balanced tree. The above approach continues to balance the structure created so far, with the phonological phrases which have not yet been joined into the structure. By doing this, the boundary values (strengths/salience) remain dependent on the values of the constituent prosodic phrases.

In the example shown in Figure 2, the left-to-right nature of application of this rule ensures that earlier material will generally be grouped lower in the structure than later material. This ties in with Gee and Grosjean's work on discourse semantics: the later in the sentence

the information, the greater the prosodic offset. It is at this stage that PP's which precede verbs are added into the structure (assuming they haven't been already). The proviso is continued that PP's should always join to the left, rather than the right. The exception to this is the PP at the end of the sentence, if there is one, which remains untouched.

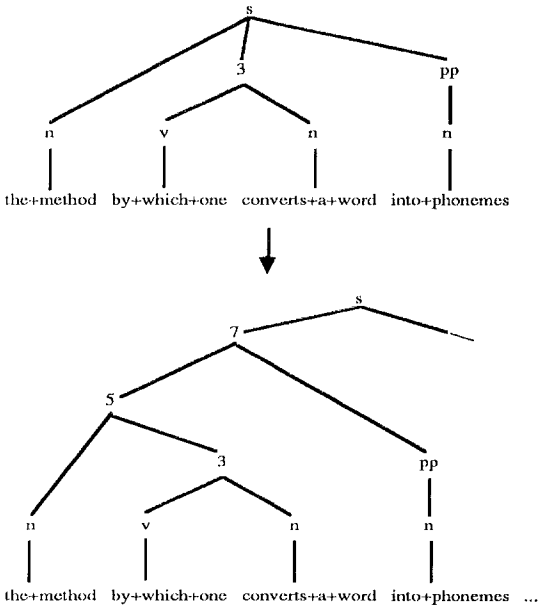


Figure 2: Verb Adjacency example

The second main change to the Bachenko and Fitzpatrick algorithm concerns boundary value assignment. Bachenko and Fitzpatrick choose to use the absolute boundary values as their reference. *Parsody* does not do this, since, according to the algorithm, the longer the sentence, the larger the values on each boundary[4] (varying from a maximum of 5 in small sentences, to 13 in the larger sample sentences). Does this mean that small sentences should have smaller boundaries, perhaps none? According to Gee and Grosjean [7; footnote 10], "It turns out (importantly) that the actual pause duration of the longest pause in each sentence does not correlate all that well (is not a factor of) the overall length of the sentence (for example, it is possible for a short, less complex sentence to have a longer main break than a longer, more complex sentence)". For this reason, in *Parsody* a normalisation algorithm is applied, so that sentences of varying lengths may have their boundaries mapped to reasonable values.

## EVALUATION

Ultimately the success of a prosody component of a TTS system will be determined by perceptual tests on the naturalness, or the acceptability, of the synthesised speech. Such tests are subjective, as well as time consuming and costly to perform, so a more objective point of reference is required.

Our approach compares the prosodic boundaries assigned by *Parsody* with data provided in Bachenko and Fitzpatrick's paper [2]. This data comprises a set of 35 sentences with the prosodic boundaries marked by hand[5]. Of these sentences, 14 are the 'original' Gee and Grosjean sentences re-analysed by Bachenko and Fitzpatrick. Our evaluation was concerned only with the primary and secondary boundaries assigned by Bachenko and Fitzpatrick.

Some points should be noted about these sentences. The Bachenko and Fitzpatrick sentences (excluding the 14 Gee and Grosjean sentences) have a fairly simple sentence structure, and should therefore be handled well by the system (*Parsody* and Bachenko and Fitzpatrick's system). In our opinion they do not constitute a rigorous test for the prosodic component of a TTS system, but they are useful for evaluation nevertheless.

The Gee and Grosjean sentences, however, have a complex sentence structure, although this is similar for each sentence. Experience would suggest that this is not a realistic sample of sentences from which to work. Bachenko and Fitzpatrick have converted these sentences to their notation. This results in each sentence having only one primary boundary, and all but one sentence having one secondary boundary. Furthermore, the primary boundary nearly always appears at the mid-point of the sentence. These results seem intuitively simple for such complex sentences, so in this evaluation the data-set "Gee and Grosjean Re-analysed" is a test against the Gee and Grosjean data, with the boundaries marked according to the normalisation algorithm employed by *Parsody*.

**Table 1 : Bachenko and Fitzpatrick Sentences**

| | *Parsody* | Bachenko & Fitzpatrick |
|---|---|---|
| Correct Primary | 21 | 16 |
| Close Primary | 8 | 9 |
| Missed Primary | 2 | 6 |
| Extra Primary | -2 | -13 |
| Primary Score | 0.778 | 0.490 |
| Correct Secondary | 13 | 11 |
| Close Secondary | 3 | 10 |
| Missed Secondary | 8 | 3 |
| Extra Secondary | 4 | 5 |
| Secondary Score | 0.583 | 0.494 |
| Overgeneration Factor | 0.965 | 0.740 |
| Overall Score | 0.681 | 0.492 |

Total Sentences - 21
Total Primary Boundaries - 31
Total Secondary Boundaries - 24
Total Tertiary Boundaries - 2

[4]The values on a boundary node can be seen in Figure 2. Values are computed by taking the sum of the phonological words at each node and adding one. A phonological word is one joined by a "+" symbol. In Figure 2, each terminal node has only one phonological word.

[5]It is important to note that even though the system assigned boundaries may be different to the human ones, the system boundaries may actually be better according to a majority expert view.

**Table 2 : Gee and Grosjean Sentences**

| | *Parsody* | Bachenko & Fitzpatrick |
|---|---|---|
| Correct Primary | 10 | 10 |
| Close Primary | 2 | 0 |
| Missed Primary | 2 | 4 |
| Extra Primary | 8 | -1 |
| Primary Score | 0.495 | 0.498 |
| Correct Secondary | 7 | 5 |
| Close Secondary | 5 | 10 |
| Missed Secondary | 3 | 0 |
| Extra Secondary | 9 | -3 |
| Secondary Score | 0.399 | 0.465 |
| Overgeneration Factor | 0.630 | 0.698 |
| Overall Score | 0.447 | 0.482 |

Total Sentences - 14
Total Primary Boundaries - 14
Total Secondary Boundaries - 15
Total Tertiary Boundaries - 31

**Table 3 : Gee and Grosjean Re-analysed Sentences**

| | *Parsody* | Bachenko & Fitzpatrick |
|---|---|---|
| Correct Primary | 10 | 10 |
| Close Primary | 3 | 1 |
| Missed Primary | 2 | 4 |
| Extra Primary | 7 | -2 |
| Primary Score | 0.700 | 0.342 |
| Correct Secondary | 6 | 7 |
| Close Secondary | 9 | 17 |
| Missed Secondary | 12 | 3 |
| Extra Secondary | -3 | -15 |
| Secondary Score | 0.355 | 0.280 |
| Overgeneration Factor | 0.913 | 0.488 |
| Overall Score | 0.528 | 0.311 |

Total Sentences - 14
Total Primary Boundaries - 15
Total Secondary Boundaries - 27
Total Tertiary Boundaries - 0

Most of the measurements given in the three tables are clear from their description. A "correct" boundary is a perfect match with the human-annotation. A "close" boundary is one where another boundary appears in its place (e.g. a secondary instead of a primary). "Extra boundary" refers to the number of boundaries produced by the system greater than the actual number of boundaries (a negative figure indicating that fewer boundaries of that type were produced).

The "scores" presented, basically provide a figure by which systems can be compared (with each other, or with human-annotated results). A score of 1 would indicate a perfect comparison of results. The figure includes both the successes and failures (including overgeneration) of the system. The overall score given, is the mean of the primary and secondary Scores.

The scores are calculated according to the following formula.

$$\text{OverGenerationFactor (OGF)} = \frac{\text{TotalBoundaries}}{\text{TotalSystemBoundaries}}$$

$$\text{Score} = \frac{(2\text{xCorrectBoundary})+\text{CloseBoundary}}{2\text{xActualBoundary}} \text{ x OGF}$$

where :

| | | |
|---|---|---|
| TotalBoundaries | = | Number of Boundaries in text |
| TotalSystemBoundaries | = | Number of Boundaries produced by system |
| CorrectBoundary | = | Number of Boundaries matched exactly |
| CloseBoundary | = | Number of boundaries matched closely (ie. a Primary marked by a Secondary, or vice versa) |
| Boundary | = | Primary or Secondary boundary |

The CorrectBoundary result is multiplied by 2 as a weighting factor. Obviously it is better to have correct boundaries than close boundaries. Accordingly, the ActualBoundary score is also doubled to maintain the scale.

Note that the *smaller* the overgeneration factor, the *larger* the amount of overgeneration (a score greater than 1 indicates undergeneration).

The results reported show that the *Parsody* system compares favourably, under this analysis, with the Bachenko and Fitzpatrick system - for example in Table 1 the overall score is 68% for Parsody, and 49% for Bachenko and Fitzpatrick's system. What is encouraging is the better performance on the prediction of primary boundaries. The automatic scoring program also presents the results in a useful way. To relate these results to Bachenko and Fitzpatrick's evaluation in [2], they quote a figure of 80%, given the assumption that primary and secondary boundaries are basically similar from a comprehensibility, and acceptability, viewpoint on the synthesised speech. This score is given by summing the Correct Primary and Close Primary scores and dividing by the total number of Primary boundaries. *Parsody* scores 93% in this case ( calculated from Table 1).

As regards our evaluation method proper, it is clear that the method requires improvement. Future methods should concentrate on punishing the incorrect placement of boundaries, especially those that affect the perception of the synthesised speech, a viewpoint that Bachenko and Fitzpatrick also seem to hold.

## CONCLUSION

This short article outlined the *Parsody* system, the essentials of which form a component of BT's Laureate Text-to-Speech system. Key features of the *Parsody* system include its ability to provide accurate parses robustly, which allows it to handle ill-formed input with ease. *Parsody* also provides a robust rule-based prosodic annotation facility, that has been developed from algorithms presented in the literature, but which have been extended for greater performance.

Most of the problems with the *Parsody* system currently lie with the parser. Despite the high performance of the word tagger, the effect of wrongly tagging a word is large, since the prosody component uses this information to construct a prosody tree in a bottom-up fashion. To improve the tagging performance we are considering including word collocation statistics. Also, it would be desirable to increase the range of syntactic structures produced by the parser. To improve the parser performance we are looking at extending the minimal grammar, but in such a way that processing speed is maintained. Future versions of the parser may also include special disambiguation rules concentrating on words having multiple pronunciations. Topic and focus marking will also be introduced at some stage.

We also hope to investigate the stochastic approach to prosodic marking. Future work will focus on assembling a suitable corpus. It is likely that the best prosodic marking procedure is one which is a hybrid of both the rule-based and stochastic-based approaches. As was mentioned earlier, the immediate goal with respect to prosodic marking has been the prediction of prosodic boundary location and of the boundary strengths. Future work will concentrate on the interpretation of boundary strengths, for example by investigating the correlation of our normalised (hence gradable) boundaries with acoustic phenomena at these boundaries.

Finally, it is important to remember the intended goal of text-to-speech systems is to synthesise unrestricted text input. Initial work has begun on extending the evaluation of the system to more 'normal' sentences. For example, work in BT's Natural Language Group includes automatic text summarisation; in tests on the summarisation of newspaper articles the length of sentences often exceeds 100 words. Our text-to-speech system must be able to handle such sentences efficiently both at the parsing and prosody stage. The lessons learnt from more difficult input such as this, may serve to increase our understanding of the relationship between syntax and prosody.


## Acknowledgements

The author would like to express his gratitude to the director of BT for permission to publish this paper.
Thanks also go to my colleagues in the Natural Language Group and the Text-to-Speech Group for their comments on this report; Keith Preston, Andy Breen, Mike Edgington, Sandra Williams, Anna Cordon and Peter Wyard. Special thanks go to Eddy Kaneen for his invaluable assistance in getting the *Parsody* system operable and for his contributions to many of the original ideas implemented in the prosodic marking component.


## REFERENCES


**[1] Allen, J.**
*Synthesis of Speech from Unrestricted Text*; Proceedings of the IEEE; Vol 4; pp. 433-442 (1976).

**[2] Bachenko, J. and Fitzpatrick, E.**
A Computational Grammar of Discourse-neutral prosodic phrasing in English.
*Computational Linguistics* Volume 16, No.3, pp.155-170 1990

**[3] Bachenko, J., and Fitzpatrick, E.**
Parsing for Prosody: What a Text-to-speech system needs from syntax
*Proceedings of IEEE Artificial Intelligence Systems in Government (AISIG)*, 1989

**[4] Bachenko, J., Fitzpatrick, E. , and Wright, C.E.**
The contribution of parsing to prosodic phrasing in an experimental text-to-speech system
*Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pp. 145-153, 1986

**[5] Cahn, J.**
From Sad to Glad: *Emotional Computer Voices*; Proceedings of Speech Tech '88; pp. 35-36 (1988).

**[6] Church K.W.**
A stochastic parts program and noun phrase parser for unrestricted text
*In Proceedings of Second Conference on Applied Natural Language Processing (ACL)*, pp. 136-143, Austin, 1988

**[7] Gee J.P. , and Grosjean F.**
Performance structures: A psycholinguistic and Linguistic Appraisal
*Cognitive Psychology*, 15: pp. 411-458, 1983

**[8] Grosjean F., Grosjean L., and Lane H.**
The patterns of silence: Performance structures in sentence production
*Cognitive Psychology*, 11: pp. 58-81, 1979

**[9] Ostendorf M., Wightman, C.W. and Veilleux, N.M.**
Parse scoring with prosodic information: An analysis / Synthesis approach
*Computer Speech and Language* (1993) 7, pp. 193-210

**[10] Selkirk, E.O.**
*Phonology and Syntax: The Relation between Sound and Structure*; MIT Press (1984).

**[11] Wang, M., and Hirschberg J.**
Predicting Intonational Boundaries Automatically from Text : the ATIS Domain
*Proceedings of the DARPA Speech and Natural Language Workshop*, Feb 1991, pp. 378-383

**[12] Wightman C.W., Veilleux N.M. , and Ostendorf M.**
Use of Prosody in syntactic disambiguation: An Analysis-by-synthesis approach
*Proceedings of the DARPA Speech and Natural Language Workshop*, Feb 1991, pp. 384-389