Assigning Prosodic Structure for Speech Synthesis: a Rule-Based Approach

Michaela Atterer

Institute of Natural Language Processing
University of Stuttgart, Germany
atterer@ims.uni-stuttgart.de

Abstract

This paper presents a model that assigns prosodic structure to unrestricted text. The model is linguistically motivated and also uses constraints on phrase length. For the implementation an XML-pipeline is used as a data-architecture. The output can be processed by a text-to-speech synthesiser for determining the locations of phrase breaks. The model outperforms another rule-based approach, and achieves either comparable results as a statistical model or comes close to those results, while being psychologically more plausible.

1. Introduction

Prosodic breaks play an important role in structuring utterances and thus increasing their understandability. In speech synthesis, good models for assigning breaks or determining prosodic structure are crucial – not only because they enhance naturalness but also because the structure these models impose serves as the input for other modules (e.g. those assigning accents or duration) and has thus a major influence on their performance.

Most text-to-speech systems employ statistical models to assign prosodic breaks to their input text [1, 2, 3]. In this work we present a rule-based approach which is more in line with work like [4, 5]. It uses insights gained from the literature in linguistics which investigates the relationship between syntactic structure and prosodic structure.

This relationship has been widely discussed. There is general agreement that – even though the existence of a syntax-prosody interface cannot be doubted – the mapping from syntax to prosody is not trivial. In general, prosodic structure is flatter than syntactic structure and less right-branching. It is usually assumed to consist of several layers (cf. Figure 3). Even though there is some debate about the number of layers and the characteristics of this hierarchy [6, 7], there seems to be a consensus that there is a layer of so-called phonological phrases just below a level of intonational phrases. The ends of intonational phrases are good places for pauses and are thus the kind of structure we aim to find in this work. Since every intonational phrase consists of a number of phonological phrases (first dubbed ϕ-phrases in [9]) we already have a good constraint for the assignment of breaks if we can find those ϕ-phrases. One of the motivations in linguistics for assuming phonological phrases is that a number of phonological rules (French liaison, English Iambic Reversal etc.) only apply within certain units [8, 9]. The major advantage of ϕ-phrases is that they can be defined syntactically, whereas intonational phrases often do not correspond to syntactic structure [8].

In this approach we use a partial parser to identify ϕ-phrases and some heuristics to find intonational phrases. We build a prosodic tree structure in XML-format which can be processed by a TTS system.

We aim at a model that is psychologically plausible, i.e. that uses linguistic constraints as far as human beings use them but also employs heuristics at a level where human beings do so. Other rule-based approaches [4, 5, 11] as well as statistical approaches pay no or little attention to the length of constituents. As length is acknowledged to be an important factor even by the linguistic literature [8], we integrate constraints on length with linguistic constraints. We also placed emphasis on enabling the model to deal with unrestricted text, which is often a problem for rule-based approaches that have to deal with parsing errors. This enables us to carry out a large-scale evaluation on a prosodically annotated corpus and to compare the model's performance with a statistical model.

2. Assignment of prosodic structure

The following is a step-by-step description of the procedure we use for assigning prosodic structure. The procedure consists of two steps: the assignment of ϕ-phrases and their bundling into intonational phrases. To obtain ϕ-phrases we modify the output of a chunk parser. To bundle them into intonational phrases we use length constraints and a balancing procedure in addition to punctuation. Technically, processing is done in a modular manner using an XML-pipeline [10] as a system architecture.

2.1. Assigning ϕ-phrases

In the literature ϕ-phrases are basically defined as consisting of a head and all its specifiers [9] or all the material to the left of the head X up to the next head outside of the maximal projection of X [8].

In our implementation ϕ-phrases are determined with the aid of Abney's chunk parser CASS [12]. Abney justifies the way his chunker works by arguing that humans have to process sentences as chunks of words. His parser imitates this by first building small units which Abney actually refers to as ϕ-phrases. These lower level chunks are defined in terms of so-called *s-heads*, which are those content words that do not stand between a function word and the content word that belongs to that function word. The latter restriction implies, inter alia, that pre-modifying adjectives do not mark chunk-boundaries. Hence *a little girl* forms a chunk, even though *little* is a content word.

Once these lower level chunks have been generated by the parser, it creates bigger units above them which correspond more to syntactic structure and are thus of less interest for the task of this work. Higher-level prosodic structure does not correspond to syntactic structure [8, 11]. We just use the ϕ-phrases to create bigger intonational units and neglect higher syntactic structure.

```
<c>
   <c0>
     <nx>
       <W C='prps'> Their</W>
       <W C='nn'> presence</W></nx>
     <vx>
       <W C='hvz'> has</W>
       <W C='vbn'> enriched</W></vx></c0></c>
 <c>
   <c0>
     <ng>
       <nx>
         <W C='dtp'> this</W>
         <W C='nn'> university</W></nx>
       <W C='cc'> and</W>
       <nx>
         <W C='dtp'> this</W>
         <W C='nn'> country</W></nx>
       <W C='cma'> ,</W>
       <W C='cc'> and</W>
       <nx>
         <W C='dtp'> many</W></nx></ng>
     <vx>
       <W C='md'> will</W>
       <W C='vb'> return</W></vx></c0>
   <nx>
     <W C='nn'> home</W></nx></c>
 <infp>
   <inf>
     <W C='to'> to</W>
     <W C='vb'> enhance</W></inf>
   <nx>
     <W C='prps'> their</W>
     <W C='jj'> own</W>
     <W C='nns'> nations</W></nx></infp>
 <W C='per'> .</W>
```

Figure 1: The structure that Abney's chunkparser CASS assigns to the sentence "Their presence has enriched this university and this country, and many will return home to enhance their own nations."

Figure 1 illustrates a piece of text that has been passed through a tagger and through CASS. The lower level-chunks are those that end with the letter *x*.

Chunks labeled *nx*, for instance, are defined as extending from the "beginning of a noun phrase to the head noun" [13]. In the sentence of Figure 1, the strings *Their presence, this university, this country, many, home* and *their own nations* belong to the category *nx*. The label *vx* basically groups auxiliaries and modals with the head verb, including all the material in between. In the sample text, *has enriched* and *will return* are examples of such verb chunks. Moreover, the text contains an infinitive chunk *inf*. We turn all these chunk categories into ϕ-phrases. Similarly, several other less frequent categories are turned into ϕ-phrases as well. They are described in more detail in [13] and [14].

If the chunks mentioned above are preceded by prepositions, complementisers, relative pronouns etc. these are included into the ϕ-phrases that are built out of the chunks to their right. This results in a structure like the one shown in Figure 2 and corresponds to the lowest levels of tree-structure shown in Figure 3. Note that the conjunction *and* has been grouped with the next *nx*.

2.2. Assigning intonational phrases

The determination of intonational phrases involves two steps. First, commas are used to mark large intonational phrases, which are then subdivided further, if they are too long. In the example, the text would be separated into two intonational phrases, where the first starts at the beginning of the sentence and ends after *country*, and the second consists of the remaining part of the sentence. To determine whether any of these intonational phrases is too long, the number of syllables of each is counted. If it exceeds a certain threshold, the intonational phrase is subdivided into constituents of roughly equal length. It goes without saying that none of the ϕ-phrases must be split by this procedure.

```
<PHI>
        <W C='prps'>Their</W>
        <W C='nn'>presence</W>   </PHI>
<PHI>
        <W C='hvz'>has</W>
        <W C='vbn'>enriched</W>  </PHI>
<PHI>
        <W C='dtp'>this</W>
        <W C='nn'>university</W>  </PHI>
<PHI>
        <W C='cc'>and</W>
        <W C='dtp'>this</W>
        <W C='nn'>country</W>  </PHI>

        <W C='cma'>,</W>
<PHI>
        <W C='cc'>and</W>
        <W C='dtp'>many</W>  </PHI>
<PHI>
        <W C='md'>will</W>
        <W C='vb'>return</W>  </PHI>
<PHI>
        <W C='nn'>home</W>  </PHI>
<PHI>
        <W C='to'>to</W>
        <W C='vb'>enhance</W>  </PHI>
<PHI>
        <W C='prps'>their</W>
        <W C='jj'>own</W>
        <W C='nns'>nations</W> </PHI>

        <W C='per'>.</W>
```

Figure 2: The same piece of text as in Figure 1. ϕ-phrases have been determined and the syntactic markup has been deleted.

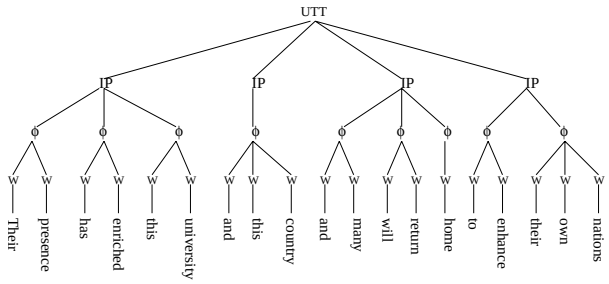

Figure 3: The prosodic structure that the model assigns to a piece of text shown in the form of a tree.

The value of the threshold accounts for the maximum number of syllables a speaker would put into one intonational phrase. Human speakers seem to have such a threshold which prevents them from forming intonational phrases that exceed a certain length [15]. The threshold can be exceeded in some cases by both the algorithm, and by a human speaker.

As this maximum number of syllables is likely to be different for different speakers, speech rates etc., it is not surprising that the optimum value for the threshold is not the same for every text. The effects of different values for the threshold can be seen in Table 1. Empirically, 13 was one of the values which tended to produce good results and was hence used as a default.

The algorithm which is in charge of dividing up large intonational phrases is shown in pseudo-code in Figure 4.

Since this algorithm always goes until the end of a ϕ-phrase when it has reached what is the optimum position for a break as far as length alone is concerned, it has a slight tendency to result in a subdivision which is not completely balanced, i.e. there is a tendency for the last I-phrase of the "new" I-phrases to become shorter than the other I-phrases. In practice, however, this seemed not to be a disadvantage. The various alternatives to this solution that were tested turned out to be inferior to this solution [14]. We believe that the reason why the present solution performs so well might be its similarity to human processing methods.

After the assignment of this higher level prosodic structure the example text looks like in Figure 5. This corresponds to

```
for each I_phrase do {
-count number_of_syllables in I_phrase
-if (number_of_syllables > 13) then {
  -number_of_breaks is number_of_syllables/13
  -optimum_length of each of the new I_phrases =
   number_of_syllables/(number_of_breaks + 1)
  -do number_of_breaks times {
    -go optimum_length syllables forward
    -go further until you reach the next end of
     a PHI-phrase
    -mark this as the end of a new I_phrase
    }
  }
}
```

Figure 4: Pseudo-code of the algorithm that assigns ϕ-phrases.

```
<UTT>
    <IP>
       <PHI>
          <W C='prps'>their</W>
          <W C='nn'>presence</W> </PHI>
       <PHI>
          <W C='hvz'>has</W>
          <W C='vbn'>enriched</W> </PHI>
       <PHI>
          <W C='dtp'>this</W>
          <W C='nn'>university</W> </PHI>
    </IP>
    <IP>
       <PHI>
          <W C='cc'>and</W>
          <W C='dtp'>this</W>
          <W C='nn'>country</W> </PHI>
    </IP>
    <IP>
       <PHI>
          <W C='cc'>and</W>
          <W C='dtp'>many</W> </PHI>
       <PHI>
          <W C='md'>will</W>
          <W C='vb'>return</W> </PHI>
       <PHI>
          <W C='nn'>home</W> </PHI>
    </IP>
    <IP>
       <PHI>
          <W C='to'>to</W>
          <W C='vb'>enhance</W> </PHI>
       <PHI>
          <W C='prps'>their</W>
          <W C='jj'>own</W>
          <W C='nns'>nations</W> </PHI>
    </IP>
</UTT>
```

Figure 5: Prosodic structure that the model assigns to a piece of text. This structure corresponds to the tree-structure in Figure 3.

the structure shown in Figure 3. To be processed by the Festival Speech Synthesis System [16], the format has to be changed slightly. These changes involve renaming the tags, adding hyperlinks etc. The basic structure, however is not affected.

The synthesiser then interprets the ends of intonational phrases as small breaks and the ends of utterances as large breaks [1].

3. Results

We evaluated the implementation of the model against both the statistical model described in [3] (which is used in Festival) and the rule-based approach by [5]. We employed the same evaluation measures that were used by [3]:

$$\textbf{Breaks-correct (BC)} = \frac{B - D}{B} \times 100\% \qquad (1)$$

$$\textbf{Junctures-correct (JC)} = \frac{N - D - I}{N} \times 100\% \qquad (2)$$

[1] This was done using a Festival version which was being developed at the University of Edinburgh at the time of this project. Processing this kind of XML-format is not yet part of the official Festival release.

text category	model	BC	JC	JI
commentary	th = 13	70.625%	89.169%	5.105%
	stm	76.875%	91.940%	3.401%
news	th = 7	61.438%	83.250%	6.868%
	th = 13	48.366%	82.915%	3.853%
	stm	59.477%	87.604%	2.010%
lecture	th = 5	75.574%	85.516%	8.404%
	th = 13	56.393%	85.475%	3.672%
	stm	64.098%	88.250%	2.815%
fiction	th = 10	72.503%	89.751%	4.257%
	th = 13	67.149%	89.719%	3.122%
	stm	75.977%	92.337%	2.428%
magazine-style	th = 7	75.124%	86.219%	7.892%
	th = 13	61.691%	87.044%	3.887%
	stm	70.647%	90.224%	2.827%
speech	th = 13	68.908%	91.117%	3.903%
	stm	73.109%	90.579%	5.114%

Table 1:
The result of the evaluation of the present model (with various parameters for the threshold th) and the statistical model (stm). Six different text categories were tested.

$$\textbf{Juncture-insertions (JI)} = \frac{I}{N} \times 100\% \qquad (3)$$

where *N, B, D* and *I* have the following meaning:

***N*:** The total number of junctures, where "juncture" is anything that is a non-break or a break; thus N corresponds to the number of whitespaces in the corpus.

***B*:** The total number of junctures which are breaks.

***D*:** Deletion error: A break is marked in the reference sentence, but not in the test sentence.

***I*:** Insertion error: A break is marked in the test sentence but not in the reference sentence.

Breaks-correct gives the percentage of breaks that have been found, and is hence comparable to the measure *recall* used in information retrieval. *Junctures-correct* gives the percentage of junctures which are assigned the same value in the test corpus as in the reference corpus. *Juncture-insertions* accounts for wrong break-assignments.

The model was evaluated using articles from the MAchine Readable Spoken English Corpus (MARSEC) [17]. This corpus is labeled with prosodic breaks. Six randomly chosen texts (belonging to six different text categories) were tested. Together they contained approximately 8000 words. For each of these texts the present model was tested with various values for the threshold mentioned above and the statistical model by [3] was tested as well. Table 1 shows the results for the optimum value of the threshold for each of the texts and for the default (in cases where the default performs worse).

On three texts the present model is slightly better as far as *Breaks-correct* is concerned. On one text it is better as far as *Juncture-insertions* is concerned. On two texts the statistical model is slightly better as far as both measures are concerned. Hence the rule-based approach seems to perform slightly worse than the statistical model. Considering, however, that the statistical model by [3] had been trained on the MARSEC corpus and thus optimizes according to this corpus we think that our rule-based account is very close if not equally good. There is a

G&G-corpus	Breaks-correct	Jct-correct	Jct-insertions
B&F	85.714%	95.906%	1.754%
present model (threshold 13)	92.857%	98.235%	0.588%
statistical model	85.714%	94.117%	3.529%

Table 2:
Evaluation of three models on the Gee&Grosjean (G&G)-corpus: Bachenko and Fitzpatrick (B&F)'s model, the present model and Taylor and Black's statistical model.

good chance that it outperforms the statistical model on a corpus consisting of data which comes from a different domain and was not seen by the latter model during training. As far as the text category speech in our test corpus is concerned, the rule-based account with threshold 13 performs very well compared to the statistical model.

In order to compare the present model with another rule-based model [5], we used the same evaluation measures on a different corpus comprising 14 sentences which had been used by [5] and [4] for evaluating their models. We also tested the statistical model on this corpus (cf. Table 2). The corpus is annotated with more than one kind of breaks, but we only considered those marked as major intra-sentential breaks.

It is quite conceivable that on this corpus the present model is not only better than the other rule-based model but also better than the statistical model. One reason for this might be that here the statistical model does not have the advantage of having been trained on the corpus.

4. Discussion

The evaluation of models that assign phrase breaks is slightly problematic. Often there is more than one way to break up a sentence into intonational phrases. The automatic evaluation method used here does not account for this fact. Other problems were that the MARSEC corpus had been used as a training corpus for the statistical model, and that all other corpora that are both suitable and available are usually relatively small and/or not unrestricted. This was also the case for the G&G-corpus. The sentences in this corpus all seemed very similar. Hence future work will evaluate the present model via a listening experiment.

Another interesting task would be a further investigation of the threshold parameter that was introduced for determining I-phrases. Table 1 suggests that it might account for effects like speech rate, speaking style or genre in some way.

5. Conclusion

We developed and implemented a model for assigning prosodic structure to unrestricted text. This model integrates linguistic and performance-based constraints in a modular manner. The implementation outputs data in XML-format which can be processed by a text-to-speech system.

The model was evaluated against a leading statistical model. Approx. 8000 words of unrestricted text were used. Even though the statistical model was trained on this corpus, the rule-based model performs similarly well. Tested on a different smaller corpus, the statistical model performed worse than the linguistically motivated model. While not being more complex than the statistical model, the present model has the advantage of being psychologically more plausible.

6. Acknowledgments

This work was carried out during the author's stay at the University of Edinburgh. I wish to thank all the people at the ICCS and CSTR who contributed to the success of the project, especially Ewan Klein for encouraging me to work on this topic and for his support.

7. References

[1] Ostendorf, M.; Veilleux, N., 1994. A hierarchical stochastic model for automatic predictions of prosodic boundary location. *Computational Linguistics*, 20(1), 27-54.

[2] Wang, M.Q.; Hirschberg, J., 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6, 175-196.

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

[4] Gee, J.P.; Grosjean, F., 1983. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15, 411-458.

[5] Bachenko, J.; Fitzpatrick, E., 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3), 155-170.

[6] Selkirk, E., 1984. *Phonology and Syntax. The relation between sound and structure.* Cambridge, MA: MIT Press.

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

[8] Nespor, M.; Vogel, I., 1986. *Prosodic Phonology.* Number 28 in Studies in Generative Grammar. Dordrecht: Foris Publications.

[9] Selkirk, E., 1981. On prosodic structure and its relation to syntactic structure. In *Nordic Prosody II: Papers from a Symposium.*, T. Fretheim (ed.). Trondheim: Tapir.

[10] McKelvie, D.; Brew, C.; Thompson, H., 1998. Using SGML as a basis for data-intensive natural language processing. *Computers and the Humanities*, 31(5), 367-388.

[11] Schweitzer A.; Haase M., 2000. Zwei Ansätze zur syntaxgesteuerten Prosodiegenerierung. In *Tagungsband der KONVENS 2000 - Sprachkommunikation*. Berlin: VDE-Verlag, 197-202.

[12] Abney, S., 1991. Parsing by chunks. In *Principle-Based Parsing: Computation and Psycholinguistics*, Berwick R.C.; Abney, S.P.; Tenny, C., eds. Kluwer.

[13] Abney, S., 1996. *Chunk stylebook*. work in progress. Available from [http://sfs.nphil.uni-tuebingen.de/ abney/].

[14] Atterer, M., 2000. *Assigning prosodic structure for speech synthesis via syntax-prosody mapping*. MSc Thesis, Division of Informatics, University of Edinburgh.

[15] Knowles, G.; Wichman, A.; Alderson, P. (eds.), 1996 *Working with Speech: Perspectives on Research into the Lancaster/IBM Spoken English Corpus*. London: Longman.

[16] Black, A. W.; Taylor, P., 1999. The festival speech synthesis system: system documentation. Available from [http://www.cstr.ed.ac.uk/projects/festival/manual/].

[17] Roach, P.; Knowles, G.; Varadi, T.; Arnfield, S., 1994. MARSEC: a machine-readable spoken english corpus. *Journal of the International Phonetic Association*, 24, 47-53.