Evaluation of PROS-3 for the assignment of prosodic structure, compared to assignment by human experts

Olga van Herwijnen, Jacques Terken
IPO, Center for User-System Interaction, Technische Universiteit Eindhoven, The Netherlands

Abstract
This paper describes the results of an evaluation of PROS-3, a system that assigns prosodic structure to text on the basis of the output of a syntactic parser. In order to evaluate the performance of PROS-3 as such and in combination with a revised algorithm for prosodic phrasing, we compare it to the prosodic structure as assigned by human experts. Also, the results of a perception experiment are presented, which show that listeners have the same preference of prosodic realization as we would expect on the basis of the comparison of the prosodic structures as assigned by PROS-3 and by human experts.

1. Introduction
The aim of the evaluation presented here is to find out whether the PROS-3 algorithm [1] assigns appropriate prosodic structure provided that the syntactic representation is correct. In this evaluation the notion of ‘prosodic structure’ refers to the distribution of phrase boundaries and accents in speech. PROS-3 implements a theory about the assignment of prosodic structure on the basis of syntactic and pragmatic information. At the same time, it constitutes a module of a system that generates prosody on the basis of syntactic information produced by a state-of-the-art syntactic parser. In the latter context, it was observed that the assignment of phrase boundaries and accents is often inadequate. However, it is unclear whether this is due to the inadequate syntactic information or to inadequacies in the theory underlying PROS-3 or both.

The evaluation described in this paper is divided into two stages: (I) the performance of PROS-3 when compared to the prosodic structure as assigned by ten experts (HUMAN representation), (II) the performance of PROS-3 compared to the prosodic structure as assigned by ten experts as adjudged in a perception experiment.

1.1. PROS-3
PROS-3 is a system that assigns prosodic structure to text on the basis of a syntactic representation of the input text. This syntactic representation describes the word category of each word together with the relations between the words.

PROS-3 contains two stages. First, a binary branching tree is computed on the basis of the syntactic representation. This metrical tree specifies the weak-strong relations between the sister nodes in the syntactic representation and which syntactic categories are eligible for Focus.

Next, the metrical tree is turned into a prosodic structure specifying the location of phrase boundaries and accents within sentences. Sentences are delimited on the basis of punctuation (period or semicolon) and sentence boundaries are realized as strong boundaries (boundary3). Within sentences, PROS-3 determines the location of boundaries of Intonational (or I-) phrases and boundaries of Phonological (or Phi-) phrases. I-phrases are application domains for rules that specify intonation and are often separated by a speech pause and marked by a pitch movement; I-phrase boundaries are realized as boundaries of the intermediate level (boundary2). Phi-phrase boundaries are application domains for supra-segmental phonological rules (e.g. they block coarticulation); Phi-phrase boundaries are realized as weak boundaries (boundary1).

The allocation of accents within phrases is based on the Focus-Accent Theory [1]. One or more constituents of a sentence are marked as +F(ocus). The relation between +F and its realization as an accent located on a word is mediated by the metrical tree: one daughter of a branching node is characterized as strong and the other as weak, depending on the functor-argument relation between the two daughter nodes. The grammar specifies which phrasal categories are eligible for focus. In PROS-3 accent is assigned to individual words [2]. The lexicon, used by PROS-3, specifies that certain words (e.g. pronouns) are typically deaccented. This may block the Focus rule from applying. In these cases, deaccentuation of words affects the strong-weak labelling. Finally, accents allocated to words within a sequence of accented words within a prosodic domain may be deleted for rhythmical reasons.

Since the metrical tree is constructed on the basis of the output of the syntactic parser, it is obvious that both phrasing and accentuation are strongly influenced by the performance of the parser.

2. Evaluation of PROS-3 compared to HUMAN representation
The first stage of the evaluation can be divided into three protocols: (A) evaluation of PROS-3 on the basis of automatically derived syntactic structure by a robust parser, (B) evaluation of PROS-3 on the basis of correct syntactic structure and (C) evaluation of PROS-3 on the basis of correct syntactic structure, in combination with a revised algorithm for prosodic phrasing.

The prosodic structure as assigned according to these three protocols will be compared to the prosodic structure as assigned by ten experts.

2.1. Method
2.1.1. HUMAN representation
Twenty sentences were selected from two Dutch newspaper articles [3]. Ten experts were asked to read these articles as if they would read the text aloud and Meanwhile decide which
words they would accentuate and where in the text they would pause (to indicate a phrase boundary). The experts (linguists or phoneticians familiar with the assignment of prosodic structure) were asked to assign a prosodic structure to the texts through annotation with markers for accents and three types of phrase boundaries (weak, intermediate and strong). From the gathered results of the ten experts a single prosodic structure was derived by computing a majority score [4]. This prosodic structure will be referred to as HUMAN representation.

Sentence (1a) shows one of the sentences that was presented to the experts and later on to PROS-3. Sentence (1b) shows an example of the HUMAN representation of the sentence. The words in capitals are the accented words, the slashes indicate the phrase boundaries (number of slashes indicates boundary level).

(1a) Hoezeer er ook een verbod geldt op het lekken uit de ministerraad, toch is het beraad niet supergeheim.

(EN) However much there is a ban on leaking information from the council of ministers, the meeting is not top secret.

(1b) HOEZEER er ook een VERBOD geldt op het LEKKEN uit de MINISTERRAAD /// TOCH is het beraad NIET SUPERGEHEIM ///

2.1.2. Evaluating the performance of PROS-3 in combination with a robust parser

Next to its quality of being a theory about prosodic structure, PROS-3 has also been implemented as a module of a system for Text-to-Speech conversion in the POLYGLOT project. In order to provide this implementation of the PROS-3 algorithm with the desired syntactic information, a robust parser, STP, was developed as part of the POLYGLOT project [5].

This parser provides a syntactic representation for every input text. Parts of the sentence that cannot be integrated into the syntactic representation are left unanalyzed and connected to the root node. The grammar rules, incorporated in the grammar, contain information about functor-argument relations between adjacent syntactic categories that is needed by PROS-3 to convert the syntactic representation into a metrical representation. In addition, the grammar rules contain information about phrasing: boundaries between major syntactic constituents are hard-coded as prosodic phrase boundaries. Elements of the sentence for which no analysis can be provided are assigned the category of major constituent, and realized as separate prosodic domains and accented.

Since STP provides the information that is needed to drive PROS-3, the combination of STP and PROS-3 was used to evaluate the performance of PROS-3 in combination with a state-of-the-art robust parser. This performance constitutes a baseline. The path described here was followed for the twenty sentences mentioned in section 2.2.1.

Sentence (1c) shows an example of the output of PROS-3.

(1c) hoezeer er ook een VERBOD geldt / op het LEKKEN uit de MINISTERRAAD // TOCH is het BERAAD / NIET / SUPERGEHEIM ///

2.1.3. Evaluation of PROS-3 on the basis of correct syntactic structure

As allocation of accents and phrase boundaries is strongly dependent on syntactic structure, we might expect that correct syntactic structure will lead to more adequate allocation of accents and phrase boundaries by PROS-3. The robust parser providing the input for the PROS-3 module doesn’t always render a proper syntactic tree. In order to test the appropriateness of the prosody assignment by PROS-3 with correct syntactic input, the syntactic tree is edited manually to obtain a proper syntactic structure.

Sentence (1d) shows an example of the output of PROS-3 with good syntactic input.

(1d) hoezeer er ook een VERBOD // geldt op het LEKKEN / uit de MINISTERRAAD // TOCH is het BERAAD / NIET / SUPERGEHEIM //

2.1.4. Evaluation of PROS-3 on the basis of correct syntactic structure, in combination with a revised algorithm for prosodic phrasing

When providing PROS-3 with adequate syntactic input, it became clear that the phrasing algorithm implemented in PROS-3 gave rather poor results. For that reason, an alternative phrasing algorithm was defined. This contains three steps:

- Step 0: Assignment of boundaries of the intermediate and strong level, based on punctuation.
- Step 1: Assignment of boundaries of the intermediate level, based on length of prosodic phrases and syntactic structure.
- Step 2: Assignment of boundaries of the weak level, based on length of prosodic phrases and syntactic structure.

The information concerning the location of phrase boundaries as determined by the algorithm is then incorporated in the syntactic representation of the sentence. This syntactic and phrasal information constitutes the input for the procedure that assigns accents to words in the sentence. The accentuation algorithm is the same as in the other protocols. Again, these procedures are performed on proper syntactic input.

Sentence (1e) shows an example of the output of the revised phrasing algorithm in combination with PROS-3 with good syntactic input.

(1e) hoezeer er ook een VERBOD geldt / op het LEKKEN uit de MINISTERRAAD // TOCH is het BERAAD NIET SUPERGEHEIM ///

2.2. Results

2.2.1. Phrase boundaries

With respect to phrasing there are two main questions. One is whether a better syntactic input leads to a better performance of the procedure for assignment of phrase boundaries. The other is whether the revised algorithm for allocation of phrase boundaries performs better than the old algorithm.

The number of boundary insertions (when a boundary is assigned by the system that has not been assigned in the HUMAN representation) and deletions (when a boundary is not assigned by the system, that has been assigned by the HUMAN representation) gives a first impression of the performance of PROS-3. However, more fine-grained measures should be applied to obtain a more revealing view on the performance. The measures computed here are precision, recall and accuracy [6] and F2-value [7]. These are measures typically used in the Information Retrieval domain.

Precision is a measure of the ratio between hits and insertions/false alarms. Recall is a measure of the ratio between hits and deletions/misses. Accuracy is the fraction of predictions
that are correct. The $F_\beta$-value is a measure used to investigate the relation between precision and recall.

Precision and recall are measures for bimodal values (zero or one; present or absent). Because there are several boundary levels, the computation of these performance measures for phrase boundaries is somewhat more difficult. We had to find a way to derive a bimodal value from the existing four-modal value for boundaries (no boundary, weak, intermediate or strong boundary). Therefore confusion matrices were computed per protocol.

To obtain a bimodal value for phrase boundaries, insertions and deletions can be computed by two methods. One method doesn’t take into account the boundary types (it only makes a distinction between boundary and no boundary). The other method does take into account boundary type (when the system assigns a lower boundary than the experts did, we call it a quasi deletion, when the system assigns a higher boundary than the experts did, we call it a quasi insertion). Quasi deletions and quasi insertions are added up to the real deletions and insertion. This second method seems more correct.

Table 1: Performance measures for phrase boundary assignment, per protocol.

<table>
<thead>
<tr>
<th></th>
<th>protocol A</th>
<th>protocol B</th>
<th>protocol C</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.31</td>
<td>0.31</td>
<td>0.53</td>
</tr>
<tr>
<td>recall</td>
<td>0.74</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.76</td>
<td>0.76</td>
<td>0.88</td>
</tr>
<tr>
<td>$F_\beta$-value</td>
<td>0.44</td>
<td>0.42</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Table 1 shows that for allocation of phrase boundaries protocol C (the protocol with the revised algorithm for assignment of phrase boundaries) performs much better than protocol A and B. This improvement applies to all four measures.

2.2.2. Accents

With respect to accentuation our main question is whether a better syntactic input leads to a better performance of the procedure for assignment of accents.

The measures mentioned in section 2.3.1 (precision, recall, accuracy and $F_\beta$-value) have also been computed for accents. Therefore, the number of insertions and deletions were computed. Table 2 gives the precision, recall and accuracy rates and the $F_\beta$-values for all three protocols described in section 2.

Table 2: Performance measures for accent assignment, per protocol.

<table>
<thead>
<tr>
<th></th>
<th>protocol A</th>
<th>protocol B</th>
<th>protocol C</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.65</td>
<td>0.71</td>
<td>0.71</td>
</tr>
<tr>
<td>recall</td>
<td>0.88</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.82</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>$F_\beta$-value</td>
<td>0.75</td>
<td>0.80</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 2 shows that for allocation of accents protocol B and protocol C perform better than protocol A. This means that PROS-3 performs better with correct syntactic input than with inaccurate syntactic structure.

2.3. Discussion

2.3.1. Phrasing

With respect to phrasing, the results of the evaluation show that there is no tendency that improved syntactic input leads to a better assignment of phrase boundaries. There is no difference between protocol A and B with respect to phrasing performance.

Furthermore, the results of the evaluation show that the revised phrasing algorithm performs remarkably better than the old phrasing algorithm. The performance measures (shown in Table 1) for protocol C are clearly higher than the performance measures for protocol A and B.

2.3.2. Accentuation

With respect to accentuation, the results of the evaluation show that correct syntactic input leads to a better assignment of accents. This tendency can be deduced from the performance measures shown in Table 2: when comparing the measures for protocol A and B, we find a somewhat better performance of protocol B than of protocol A.

Furthermore, the results of the evaluation show that correct assignment of accents is not dependent on the revised phrasing algorithm. This seems to be due to the bottom-up approach for accentuation. When we compare the performance measures for protocol B versus protocol C (shown in Table 2), we see that there is no noticeable difference between protocol B versus protocol C.

In summary, when assigning prosodic structure with PROS-3 we get the best result if we apply PROS-3 in combination with the revised phrasing algorithm on correct syntactic input. For all protocols there is a certain amount of discrepancy between the HUMAN representation and automatically generated prosody. However, impressionistically, the prosodic phrasing provided by protocol A and B is often inadequate, while that for protocol C is usually appropriate. With respect to accentuation, providing PROS-3 with correct syntactic information already gives a considerable improvement, but this can only be appreciated in combination with adequate phrasing. A perception experiment was performed to put to test this impression.

3. Perception experiment

Ostendorf and Veilleux [8] already suggested that "the best test of a phrase break algorithm is in perceptual judgements of synthesized speech", in our opinion this holds not only for prosodic phrasing but same is true for accentuation. Therefore, the perception experiment presented here puts to test the results of the comparison between the assignment of prosodic structure by the three protocols and the HUMAN representation. In this experiment listeners judgements are collected about acceptability of the prosodic structure.

3.1. Method

3.1.1. Material

The twenty sentences mentioned in section 2.2.1 were processed by the IPO TTS system. Grapheme input was processed by this system, resulting in a phoneme representation, which was corrected manually. Furthermore, the prosodic structures resulting from the three protocols and the HUMAN representation were assigned. For each sentence four versions were generated (versions with the prosodic structure perceived by protocol A, B and C and the HUMAN representation).
3.1.2. Approach
The sentences are presented pair-wise to 20 listeners. These pairs consist of two versions of one sentence. Pairs are presented in all possible sequences (20*4*3=240 pairs) The listeners were asked to indicate which of the two sentences was the most acceptable by clicking with the mouse on a button on the screen. All listeners were students from Eindhoven University of Technology, with no hearing impairment. The sentences were presented over headphones. Because of the duration of the experiment we presented only half of the stimuli (120) to each listener (partitioning by latin-square).

3.2. Results
Figure 1 shows to what extent each version is preferred by the listeners. When the listeners have no preference for one of the versions, all versions would be rated for 25%.

The values in Figure 1 indicate that version B is not preferred above version A, that version C and H are highly preferred above version A and B, and that the listeners slightly prefer version H (HUMAN reference) above version C.

Figure 1: Extent of preference per version.

A more detailed comparison is made in Table 3. This table shows exactly to what extent listeners prefer version C and H above version A and B, to what extent version H is preferred above version C and that listeners have no preference for version A above B or vice versa. A score of 70.7 corresponds to a d' of 1, being a significant difference at p = .05.

Table 3: Listeners preferences when comparing two versions.

<table>
<thead>
<tr>
<th>version 1</th>
<th>version 2</th>
<th>preferred 1</th>
<th>preferred 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>0.49</td>
<td>0.51</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>0.30</td>
<td>0.70</td>
</tr>
<tr>
<td>A</td>
<td>H</td>
<td>0.26</td>
<td>0.74</td>
</tr>
<tr>
<td>B</td>
<td>C</td>
<td>0.31</td>
<td>0.69</td>
</tr>
<tr>
<td>B</td>
<td>H</td>
<td>0.26</td>
<td>0.74</td>
</tr>
<tr>
<td>C</td>
<td>H</td>
<td>0.42</td>
<td>0.58</td>
</tr>
</tbody>
</table>

3.3. Discussion
The results of the perception experiment show that listeners have no preference for sentences generated on the basis of PROS-3 with automatically derived syntactic structure or sentences generated on the basis PROS-3 with correct syntactic structure. From this we can conclude that improved syntactic structure alone doesn't improve the acceptability of the prosodic structure, according to the listeners' judgements. However, the difference between these two versions and the version of the sentences based on correct syntactic structure in combination with the revised algorithm for prosodic phrasing, led to a major preference for the latter version above the others.

Still, listeners have a slight preference for the HUMAN reference version (version H) above the most acceptable algorithm-based version (version C).

4. Conclusion
From the results of the objective evaluation of the performance of PROS-3 when compared to the HUMAN reference we conclude that we get the best result if we apply PROS-3 in combination with the revised algorithm for prosodic phrasing on correct syntactic input. Providing PROS-3 with correct syntactic information already gives a considerable improvement with respect to accentuation, but we expect that this can only be fully appreciated in combination with adequate phrasing.

The results of the perception experiment show that indeed there is no preference for the version based on correct syntactic structure above the version based on automatically generated syntactic structure, and that listeners highly prefer the version based on correct syntactic structure in combination with the revised algorithm for prosodic phrasing, above the other two algorithm-based versions. These results support our expectation that improved accentuation can only be appreciated in combination with adequate phrasing.

5. References