Do speakers realize the prosodic structure they say they do?

Olga van Herwijnen, Jacques Terken

IPO, Center for User-System Interaction, Technische Universiteit Eindhoven, The Netherlands

o.m.v.herwijnen@tue.nl

Abstract

In this paper we describe a study in which a comparison was made between prosodic structures as realized in a spoken version of a text and as assigned by annotators of this text on paper. The prosodic structures were assigned by experts. This study puts to test the strategy of annotating text on paper to obtain a HUMAN reference of the prosodic structure that would be assigned when reading text aloud. This strategy is less time consuming than the often used analysis of spoken versions to obtain the assigned prosodic structure. The results of the comparison described here show that speakers are fairly capable of predicting what prosodic structure they would assign when reading text aloud.

1. Introduction

Many evaluation studies of automatic assignment of prosodic structure to text make use of a HUMAN reference [1], i.e. a prosodic structure obtained from annotators who annotated a text with markers for accents and phrase boundaries. In these studies annotators are asked to indicate which words would be accented when they read a text aloud, and on which junctures phrase boundaries would be realized. In such studies the assumption is made that the annotators are able to accomplish this task well and that the prosodic structure indicated by the annotators corresponds to the prosodic structure of the text when they are actually reading the text aloud.

Another strategy corresponds to the prosodic structure of the text when they are actually reading the text aloud. This strategy is far more time consuming we investigate the correspondence between the two types of reference. That is, one reference is obtained from the experts prosodic structures which they predict they would assign when reading text aloud.

2. HUMAN reference

2.1. Method

Twenty sentences were selected from two Dutch newspaper articles [4]. Ten experts (linguists or phoneticians familiar with the assignment of prosodic structure) were asked to read these articles as if they would read the text aloud and meanwhile decide which words they would accented and where in the text they would pause (to indicate a phrase boundary). The experts were asked to assign a prosodic structure to the texts through annotation with markers for accents and phrase boundaries (four levels of boundaries were distinguished).

From the gathered results of the ten experts a single prosodic structure was derived. For each word and each juncture (i.e. a potential location for a phrase boundary) we decided whether there should be an accent or a phrase boundary respectively, according to the following procedures.

Distribution of accents: For every word in the text the annotations of the ten experts were summed. Only when seven or more experts marked a word for accent, then the word would be accented in the HUMAN reference, otherwise the word would remain unaccented.

A disadvantage of this method is that for a word sequence a b c various accent patterns are possible. When part of the experts assigned one pattern and part of the experts assigned another pattern, the consequence might be that in the HUMAN reference an unintentional pattern was derived (i.e. a pattern that has not been assigned by any of the experts). However, since alternative methods are much more complicated and may impose other disadvantages, we decided to use the straightforward method outlined above. In most cases it turned out to be unproblematic.

Distribution of phrase boundaries: Four levels of phrase boundary were distinguished: no boundary (level 0), weak boundary (level 1), intermediate boundary (level 2) and strong boundary (level 3). For every juncture in the text, the mean score of the experts was computed by summing the scores of the ten experts (Σmax = 30). The summed scores were mapped onto boundary values for each juncture in the following way.

<table>
<thead>
<tr>
<th>Score</th>
<th>Boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 6</td>
<td>no boundary</td>
</tr>
<tr>
<td>7 - 14</td>
<td>weak boundary</td>
</tr>
<tr>
<td>15 - 23</td>
<td>intermediate boundary</td>
</tr>
<tr>
<td>24 - 30</td>
<td>strong boundary</td>
</tr>
</tbody>
</table>

2.2. Expert agreement

In order to be allowed to compute a mean text representation from the results of the experts, there should be a reasonable level of agreement between experts. To assess the agreement on classification tasks, the kappa coefficient [5] is used. The kappa coefficient \( K \) measures pairwise agreement among a set of coders making category judgements, correcting for expected chance agreement.

\[
K = \frac{(P(A) - P(E))}{(1 - P(E))}
\]

(1)

Here \( P(A) \) is the proportion of times the coders agree and \( P(E) \) is the proportion of times that the coders are expected to agree by chance.

For allocation of accents \( K = 0.72 \), for allocation of phrase boundaries \( K = 0.66 \). Content analysis researchers generally think of \( K > 0.8 \) as good reliability, with \( 0.67 > K > 0.8 \) allowing tentative conclusions to be drawn [5]. From this it seems...
that we have only modest reliability. However, there were clear
differences in the numbers of accents and phrase boundaries al-
located by the individual experts (see Table 1). We assume that
the relatively modest kappa’s are mainly the consequence of this
variation in the amount of accents and boundaries assigned by
the experts. Support for this interpretation comes from the ob-
servation that for any two experts who have approximately the
same amount of accents and boundaries, pairwise kappa’s are
around 0.78. Therefore, we consider it valid to compute a sin-
gle prosodic structure on the basis of the expert annotations.
This ‘mean’ prosodic structure will be referred to as HUMAN
reference.

For comparison, the agreement between three Text-to-
Speech systems for Dutch and the HUMAN reference is some-
what lower, especially for accents [6]. There, $K = 0.53$ for allo-
cation of accents and $K = 0.65$ for allocation of phrase bound-
aries.

Table 1: Number of accents and phrase boundaries per expert and HUMAN reference (paper version).

<table>
<thead>
<tr>
<th>expert</th>
<th># accents</th>
<th># bound3</th>
<th># bound2</th>
<th># bound1</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>75</td>
<td>33</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>02</td>
<td>136</td>
<td>21</td>
<td>15</td>
<td>55</td>
</tr>
<tr>
<td>03</td>
<td>151</td>
<td>20</td>
<td>19</td>
<td>27</td>
</tr>
<tr>
<td>04</td>
<td>136</td>
<td>20</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>05</td>
<td>129</td>
<td>21</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>06</td>
<td>122</td>
<td>21</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>07</td>
<td>114</td>
<td>29</td>
<td>21</td>
<td>26</td>
</tr>
<tr>
<td>08</td>
<td>138</td>
<td>20</td>
<td>17</td>
<td>28</td>
</tr>
<tr>
<td>09</td>
<td>105</td>
<td>20</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>130</td>
<td>20</td>
<td>20</td>
<td>22</td>
</tr>
<tr>
<td>HUMAN</td>
<td>112</td>
<td>20</td>
<td>17</td>
<td>16</td>
</tr>
</tbody>
</table>

3. Comparison

The comparison of the prosodic structures of spoken versions
of the sentences with those of the paper versions consists of two
parts. First, the spoken versions are compared to the speakers’
own paper versions. Second, a SPOKEN reference is computed
and compared to the HUMAN reference.

3.1. Production experiment

A production experiment was performed to obtain the spoken
versions from the text. Three of the ten experts mentioned above
were asked to read the two newspaper articles [4] aloud. From
these spoken versions of the texts, the same twenty selected
sentences were taken. We analyzed these sentences to obtain
the prosodic structures that the speakers realized. The analysis
consisted of two parts: (I) pitch contour analysis together with
auditory analysis to indicate which words were accented, and
(II) analysis in the time frame together with auditory analysis
to indicate on which junctures phrase boundaries were realized.
The level of the phrase boundary was determined on the basis of
pause duration and segmental factors such as word final length-
ening.

As mentioned in section 2.2 there should be a reasonable
level of agreement between experts to be allowed to compute
a mean representation. To assess this agreement, we computed
the kappa coefficient again. For allocation of accents $K = 0.77$,
for allocation of phrase boundaries $K = 0.73$. This means that
the agreement between the three experts was rather high (see
section 2.2). Therefore we consider it valid to compute the SPO-
KEN reference.

The distribution of boundaries corresponds with that de-
scribed in section 2.1. However, as the spoken versions are pro-
duced by only three experts, we had to adjust the criteria
for accent or no accent and for phrase boundary level. In this
SPOKEN reference a word is accented when two or three of the
speakers assigned accent to that word. Phrasing is somewhat
more difficult again, because of the four levels of phrase bound-
aries (no boundary, weak boundary, intermediate boundary and
strong boundary). The criteria for distribution of phrase bound-
aries are given below.

| 0 - 1 | no boundary |
| 2 - 4 | weak boundary |
| 5 - 7 | intermediate boundary |
| 8 - 9 | strong boundary |

3.2. Spoken version versus paper version

For all three speakers the prosodic structures of their spoken
versions of the twenty sentences were compared to the prosodic
structures of their own paper version of the same twenty sen-
tences. Table 2 shows the numbers of accents and phrase bound-
aries assigned in the spoken version for all three experts and for
the SPOKEN reference.

Table 2: Number of accents and phrase boundaries per expert and SPOKEN reference (spoken version).

<table>
<thead>
<tr>
<th>expert</th>
<th># accents</th>
<th># bound3</th>
<th># bound2</th>
<th># bound1</th>
</tr>
</thead>
<tbody>
<tr>
<td>04</td>
<td>130</td>
<td>21</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>05</td>
<td>129</td>
<td>21</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>08</td>
<td>133</td>
<td>21</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>SPOKEN</td>
<td></td>
<td>131</td>
<td>13</td>
<td>16</td>
</tr>
</tbody>
</table>

When we compare these numbers to those for the paper ver-
(see Table 1), we see that there is no significant difference
in number of accents and phrase boundaries between the spoken
version and the paper version of all three experts. We also see
that there are no extreme differences in the number of accents
and phrase boundaries assigned by the three speakers. These
results give a first impression of the capability of speakers to pre-
dict on paper which prosodic structure they would assign when
they actually read the text aloud.

To obtain a more revealing view on the performance more
fine-grained measures should be applied. The measures com-
puted here are precision, recall and accuracy [7]. These are
measures typically used in the Information Retrieval domain.

**Precision** is a measure of the ratio between hits and inser-
tions/false alarms. **Recall** is a measure of the ratio between hits
and deletions/misses. **Accuracy** is the fraction of predictions
that are correct.

Table 3: Computation of Precision, Recall and Accuracy.

<table>
<thead>
<tr>
<th>reference</th>
<th>accent</th>
<th>no accent</th>
</tr>
</thead>
<tbody>
<tr>
<td>test case</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>
Table 3 and equations 2-4 show how the measures precision, recall and accuracy are computed for accents. For phrase boundaries the performance measures are computed in a similar way.

\[
\text{precision} = \frac{A}{A + B} \quad (2)
\]

\[
\text{recall} = \frac{A}{A + C} \quad (3)
\]

\[
\text{accuracy} = \frac{A + D}{A + B + C + D} \quad (4)
\]

In these equations, \(B\) denotes insertions and \(C\) denotes deletions. The precision rate becomes higher as the number of insertions decreases. The recall rate becomes higher as the number of deletions decreases.

A measure to investigate the relation between precision and recall is the \(F_\beta\)-value [8]. This measure can be computed with equation 5.

\[
F_\beta = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}} (5)
\]

When \(\beta = 1\) the precision rate and recall rate have the same weights. When \(\beta\) is chosen zero, then the \(F_0\)-value equals the precision rate. Because precision and recall rate are of equal interest, the assumption \(\beta = 1\) was made.

### 3.2.1. Accents

Table 4 gives these performance measures for accents for the three speakers. Here, the paper version was taken as reference and the spoken version was taken as test case.

Table 4: Precision, recall, accuracy and \(F_\beta\)-value for accent assignment per expert.

<table>
<thead>
<tr>
<th></th>
<th>E04</th>
<th>E05</th>
<th>E08</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.89</td>
<td>0.78</td>
<td>0.89</td>
</tr>
<tr>
<td>recall</td>
<td>0.85</td>
<td>0.78</td>
<td>0.86</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.91</td>
<td>0.84</td>
<td>0.91</td>
</tr>
<tr>
<td>(F_\beta)-value</td>
<td>0.87</td>
<td>0.78</td>
<td>0.88</td>
</tr>
</tbody>
</table>

With respect to accentuation, these results show that the spoken versions of the sentences correspond rather good with the paper versions of the sentences. The performance measures for expert 05 are somewhat lower than those for expert 04 and expert 08, but are still reasonably good. This means that speakers are capable of predicting which words they would accentuate when reading text aloud.

### 3.2.2. Phrase boundaries

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 expert.

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). However, there is a large difference between strong boundaries and weak boundaries. For phrase boundaries a standard consistency criterion is agreement within +/- 1 level [9]. Therefore, we used a second (rather stringent) method that 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.

We computed the performance measures by both methods. The first method shows to what extent speakers are able to predict where the would produce a phrase boundary. The second method is even more exact, it shows to what extent speakers are able to predict where they would produce a phrase boundary and of which level that boundary will be.

Table 5 gives the performance measures for phrase boundaries for the three speakers. Again, the paper version was taken as reference and the spoken version was taken as test case.

Table 5: Precision, recall, accuracy and \(F_\beta\)-value for phrase boundary assignment per expert.

<table>
<thead>
<tr>
<th></th>
<th>E04</th>
<th>E05</th>
<th>E08</th>
<th>E04</th>
<th>E05</th>
<th>E08</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.81</td>
<td>0.73</td>
<td>0.91</td>
<td>0.66</td>
<td>0.73</td>
<td>0.88</td>
</tr>
<tr>
<td>recall</td>
<td>0.85</td>
<td>0.84</td>
<td>0.77</td>
<td>0.79</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.96</td>
<td>0.94</td>
<td>0.94</td>
<td>0.93</td>
<td>0.92</td>
<td>0.93</td>
</tr>
<tr>
<td>(F_\beta)-value</td>
<td>0.83</td>
<td>0.78</td>
<td>0.83</td>
<td>0.72</td>
<td>0.67</td>
<td>0.77</td>
</tr>
</tbody>
</table>

When we consider method 1, with respect to phrasing, these results show that the spoken versions of the sentences correspond rather good with the paper versions of the sentences. The performance measures for expert 05 are somewhat lower than those for expert 04 and expert 08, but are still reasonably good. This means that speakers are capable of predicting where they would allocate phrase boundaries when reading text aloud.

When we consider method 2, the performance measures are somewhat less promising. The measures for expert 08 are still reasonably good, but the measures for expert 04 and expert 05 are worse. This means that though speakers are capable of predicting where they would allocate phrase boundaries, they are somewhat less capable of predicting the level of those allocated boundaries.

### 3.3. SPOKEN reference versus HUMAN reference

First, the number of accents and phrase boundaries assigned by the HUMAN reference and the SPOKEN reference were compared. There is no significant difference in the number of phrase boundaries assigned by both references. However, the number of accents assigned by the SPOKEN reference is somewhat higher than the number of accents assigned by the HUMAN reference.

To obtain a more revealing view on the performance, again the precision, recall, accuracy and \(F_\beta\)-value were computed. Table 6 gives the performance measures for accents and phrase boundaries (for both methods described in section 3.2.2). Here, the HUMAN reference was taken as reference and the SPOKEN reference was taken as test case.

These results show that with respect to accentuation the 'mean' spoken version corresponds rather good with the 'mean' paper version. With respect to phrasing, the SPOKEN reference
Table 6: Precision, recall, accuracy and $F_\beta$-values for comparison between HUMAN reference and SPOKEN reference.

<table>
<thead>
<tr>
<th></th>
<th>method 1 bound</th>
<th>method 2 bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>0.81</td>
<td>0.92</td>
</tr>
<tr>
<td>recall</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>accuracy</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td>$F_\beta$-value</td>
<td>0.87</td>
<td>0.89</td>
</tr>
</tbody>
</table>

corresponds rather good with the HUMAN reference when we consider method 1. When we consider method 2, the correspondence is slightly lower, but still rather good.

4. Discussion

The results of the experiment described here show that speakers are rather well capable of predicting what prosodic structure they would assign to text when reading the text aloud. This follows from the reasonably high precision, recall and accuracy rates and $F_\beta$-values that were computed to test the agreement between paper versions and spoken version from the experts. The same is true for the agreement between the HUMAN reference and the SPOKEN reference.

One remark should be made. Speakers are rather well capable of predicting which words would be accented when reading a text aloud and they are rather well capable of predicting on which junctures a phrase boundary would be realized. However, speakers are somewhat less capable of predicting what level phrase boundary would be realized on that specific juncture.

Still there is some variation between the paper version and spoken version for each single expert. This variation can be partially explained by the fact that there was a time span between the annotation on paper and the production of the spoken version. We expect that the agreement would be even higher when the recordings would have been made right after the annotation on paper. Another explanation is that there is some variation possible in the assignment of prosodic structure. When speakers pronounce the same sentence a number of times, they will not always realize the same prosodic structure. This means that when we compare two spoken versions of the same sentence, produced by the same speaker, we would also find some variation in prosodic structure.

5. Conclusion

From the results of the comparison between the paper versions and the spoken versions for all three experts and the comparison between the HUMAN reference and the SPOKEN reference we conclude that speakers are capable of predicting what prosodic structure they will realize when reading text aloud. This means that annotating text on paper is an acceptable strategy to obtain the prosodic structure that would be realized when reading the text aloud. This strategy can be used instead of the more time consuming strategy where the prosodic structure is obtained by analysis of spoken text.

6. References