Speech Prosody 2016 poster (A4)

Dataset · May 2016

3 authors:

Mireia Farrús
University Pompeu Fabra
40 PUBLICATIONS  169 CITATIONS
SEE PROFILE

Catherine Lai
The University of Edinburgh
26 PUBLICATIONS  120 CITATIONS
SEE PROFILE

Johanna D Moore
The University of Edinburgh
227 PUBLICATIONS  5,598 CITATIONS
SEE PROFILE
Paragraph-based Prosodic Cues for Speech Synthesis Applications

Mireia Farrús1, Catherine Lai2, Johanna D. Moore2
1TALN Research Centre, Universitat Pompeu Fabra, Barcelona
2School of Informatics, University of Edinburgh, UK

Abstract

Obtaining full expressiveness in multi-sentential synthesized discourse is a challenge, since the prosodic differences observed in discourse units are usually not taken into account. We analyze the prosody of paragraph units in a corpus of TED Talks and perform classification experiments in order to identify the most consistent features used to distinguish paragraph breaks. The results show that: (1) differences in prosody related to paragraph position are significant, and (2) boundary features are the most consistent cues in marking paragraph boundaries. This suggests that these features should be taken into account when generating spoken discourse in order to improve naturalness and expressiveness.

Motivation & Objectives

Currently... TTS naturalness and expressiveness are improved by increasing the prosodic variability based on inferred affective states.

However... a more discourse structure aware approach is needed to improve speech synthesis naturalness.

Therefore, in this work, we...

(1) Confirm the presence of prosodic cues for paragraph structure in a large, varied corpus of semi-spontaneous speech.
(2) Analyze prosodic patterns in terms of F0, intensity and timing of sentences with respect to paragraph position and boundaries.
(3) Perform SVM classification experiments to investigate how consistently and robustly these features appear at paragraph boundaries.

Feature Statistical Analysis

• Mean values for sentence-based features in different paragraph positions:

<table>
<thead>
<tr>
<th></th>
<th>First</th>
<th>Middle</th>
<th>Last</th>
</tr>
</thead>
<tbody>
<tr>
<td>max(norm.F0)</td>
<td>8.43</td>
<td>7.59</td>
<td>7.40</td>
</tr>
<tr>
<td>mean(norm.F0)</td>
<td>-1.87</td>
<td>-1.86</td>
<td>-1.24</td>
</tr>
<tr>
<td>min(norm.F0)</td>
<td>-10.49</td>
<td>-11.35</td>
<td>-11.63</td>
</tr>
<tr>
<td>max(norm.I)</td>
<td>4.99</td>
<td>5.18</td>
<td>5.06</td>
</tr>
<tr>
<td>mean(norm.I)</td>
<td>2.37</td>
<td>2.09</td>
<td>2.98</td>
</tr>
<tr>
<td>min(norm.I)</td>
<td>-6.28</td>
<td>-5.12</td>
<td>-5.78</td>
</tr>
</tbody>
</table>

• Recurrent patterns:

(a) decrease through the paragraph: all F0 and I features F0 (except for median.norm.I)
(b) reduced values in the middle position: variation-related features: range, sd, abs(fldiff), abs(slope)
(c) increase in the middle position: spk.rate, median.norm.I

• Mean values for boundary-based features in initial/middle (no break) versus final (break) position:

<table>
<thead>
<tr>
<th></th>
<th>no break</th>
<th>break</th>
</tr>
</thead>
<tbody>
<tr>
<td>fldiff.F0</td>
<td>-0.22</td>
<td>0.91</td>
</tr>
<tr>
<td>fldiff.I</td>
<td>-0.13</td>
<td>1.45</td>
</tr>
<tr>
<td>abs(fldiff.I)</td>
<td>2.73</td>
<td>3.88</td>
</tr>
<tr>
<td>pause.dur</td>
<td>0.64</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Conclusions

• There is clear evidence of prosodic resets over paragraph breaks
• We can also observe a steady declination in prosodic level over the paragraph
• Difference features are more discriminative of boundaries than sentence-based features
• Paragraphs have an identifiable suprasentential prosodic structure that can be described in terms of relative changes in F0, intensity, and timing
• The classification experiments support the idea that utterance intrinsic features to paragraph position exist
• Pause duration is the most robust predictor of paragraph breaks

We should be able to employ paragraph declination, pause and prosodic reset features to improve the naturalness of longer synthesized speech

Experimental Setup

Data. 1365 TED talks – 1156 different English speakers, various accents.

Sentence Alignment. Word timings obtained through Viterbi forced alignment (ASR system), used later to obtain sentence boundaries.

Prosodic Features. Normalized F0 & Intensity contours (Praat).

• Aggregate statistics: mean, max, min, median, slope
• Differences between: first–last words of a sentence (fldiff) 99th and 1st quartile (range) target and following sentence (sdiff) last word of target sent. and next (lndiff)
• Speaking rate (spk.rate), pause duration (pause.dur)

Classification. SVM classification experiments

• 10-fold cross-validation (LibSVM – Weka)
• C-SVC approach, RBF kernel

Baseline accuracy: 50% (break/no break classes equally balanced)

Classification Experiments

ROC curve for the 2 best feature set (pause.dur + sdiff.norm.F0), break class.