TOWARD THE USE OF INFORMATION DENSITY BASED DESCRIPTIVE FEATURES IN HMM BASED SPEECH SYNTHESIS

Sébastien Le Maugér¹, Bernd Möbius¹, Ingmar Steiner¹,²

¹ Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany, ²DFKI

Introduction

- Background
  - Statistical TTS = huge effort assigned to acoustic modelling
  - Descriptive feature set = almost the same for each system (the one presented in [1])
- Problem
  - How to enrich this descriptive feature set?
- Proposition
  - New descriptive feature = unpredictability of an event
  - Based on information density & widely used in computational linguistics

Unpredictability (Surpasil)

\[ \text{Surpasil}(U_i) = -\log P(U_i|U_{i-1}, U_{i-1-1}, \ldots) \]

- Predictability of a word correlates with processing effort of pronouncing this word [2]
- Same correlation found at the syllable level [3]

Feature Generation

- Syllable based
  - IPA phoneme representation
- Word based
  - All punctuation marks are discarded
  - A break mark is inserted at the end of each paragraph
  - All words are converted to lower case

Objective Evaluation

- Speech corpus
  - From “Black Beauty” (2013 Blizzard Challenge)
  - 1 h (~470 utterances) = 13 522 syl., 7038 words
  - Segmented using EMHMM + manually corrected
- Text corpus
  - 2013 Blizzard Challenge – “Black Beauty”
  - 82 books = 951 316 syl., 1 973 368 words
- System setup
  - HTS 2.2 standard configuration
  - Vocoder = STRAIGHT + MLSA filter
  - MGC (50) + LF0 (1) + BAP (25) + \( \Delta + \Delta \Delta \)
  - Word-prosody
  - Phrase-prosody

Distance Analysis

<table>
<thead>
<tr>
<th>Condition</th>
<th>MCD</th>
<th>RMS-F0</th>
<th>VER</th>
<th>RMS-dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline unpred</td>
<td>6.45</td>
<td>475</td>
<td>15</td>
<td>11.1</td>
</tr>
<tr>
<td>unpred_syllable</td>
<td>6.33</td>
<td>463</td>
<td>14.6</td>
<td>10.6</td>
</tr>
<tr>
<td>unpred_all</td>
<td>6.33</td>
<td>467</td>
<td>14.8</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Tree Analysis

<table>
<thead>
<tr>
<th>categories</th>
<th>baseline</th>
<th>unpred_syl</th>
<th>unpred_all</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>1648</td>
<td>1615</td>
<td>1594</td>
</tr>
<tr>
<td>p3</td>
<td>2134</td>
<td>2037</td>
<td>1175</td>
</tr>
<tr>
<td>p5</td>
<td>0</td>
<td>0</td>
<td>694</td>
</tr>
<tr>
<td>syl-position</td>
<td>5799</td>
<td>5652</td>
<td>4056</td>
</tr>
<tr>
<td>word-position</td>
<td>7836</td>
<td>6757</td>
<td>4202</td>
</tr>
<tr>
<td>word-prosody</td>
<td>2028</td>
<td>2817</td>
<td>2198</td>
</tr>
<tr>
<td>phrase-position</td>
<td>1184</td>
<td>1573</td>
<td>802</td>
</tr>
<tr>
<td>phrase-prosody</td>
<td>8723</td>
<td>8323</td>
<td>5892</td>
</tr>
<tr>
<td>utterance</td>
<td>7260</td>
<td>7429</td>
<td>6799</td>
</tr>
</tbody>
</table>

Subjective Evaluation

<table>
<thead>
<tr>
<th>Preference Evaluation</th>
<th>Evaluated system</th>
<th>baseline</th>
<th>unpred_all</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>72.6</td>
<td>27.4</td>
</tr>
</tbody>
</table>

Analysis

- AB
  - Clear preference for the proposed system
- MUSHRA
  - Improvement \( \Rightarrow \) just a tendency
- Evaluation: Spectrum vs. prosody?
- Global
  - Assumption = spectrum not impacted, prosody + natural

Conclusion

- New descriptive feature: unpredictability (widely used in computational linguistics)
- Full process to compute and apply these features
- Objective analysis
  - Similarity not impacted
  - Model uses this feature into account
- Subjective evaluation: preference for our system \( \Rightarrow \) which dimension?

Bibliography