The PAVOQUE corpus as a resource for analysis and synthesis of expressive speech

Ingmar Steiner1,2,3, Marc Schröder3, Annette Klepp1,3
steiner@coli.uni-saarland.de

1Multimodal Computing & Interaction
2Saarland University
3German Research Center for Artificial Intelligence

Overview

We announce the release of the PAVOQUE corpus, a single-speaker, multi-style database of German speech, designed for analysis and synthesis of expressive speech. The corpus has been previously used for voice conversion [5] and expressive text-to-speech synthesis [1, 4]. The full corpus data is now being made available to the public, under a Creative Commons license. It is hosted at https://github.com/marytts/pavoque-data.

Speaker and recordings

Stefan Röttig, a male native speaker of German trained as a professional actor and baritone opera singer, was hired to produce the corpus. The recordings were carried out in a sound-proof studio, over multiple sessions, with a sampling rate of 44.1 kHz at 24 bit per sample. All utterances were automatically transcribed using MaryTTS [3]; the phone-level segmentation was manually verified by phonetically trained research assistants.

Corpus composition

The bulk of the corpus consists of general-domain sentences (A) automatically extracted from Wikipedia using a greedy algorithm optimizing for phonetic and prosodic coverage [2]; these were spoken in a neutral, “news-reading” style. 375 more of these (B) are common to all styles. A number of domain-specific utterances (C) were spoken as well.

<table>
<thead>
<tr>
<th>set</th>
<th>neutral</th>
<th>cheerful</th>
<th>depressed</th>
<th>aggressive</th>
<th>poker</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2639</td>
<td>25</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>B</td>
<td>375</td>
<td>375</td>
<td>375</td>
<td>375</td>
<td>375</td>
</tr>
<tr>
<td>C</td>
<td>112</td>
<td>184</td>
<td>156</td>
<td>201</td>
<td>175</td>
</tr>
<tr>
<td>total</td>
<td>3126</td>
<td>584</td>
<td>556</td>
<td>601</td>
<td>575</td>
</tr>
<tr>
<td>time</td>
<td>321 min</td>
<td>46 min</td>
<td>55 min</td>
<td>45 min</td>
<td>49 min</td>
</tr>
</tbody>
</table>

Overall, 5442 utterances (8 h 37 min) are available in five different speaking styles.

Selected statistics

- Articulation rate (phones/s)
- F0 (Hz)
- Spectral slope (dB)

References