Assessing a Speaker for Fast Speech in Unit Selection Speech Synthesis

Donata Moers1,2, Petra Wagner2, Bernd Möbius1,3, Igor Jauk2, Filip Müllers1
1Department of Speech and Communication, University of Bonn;
2Faculty of Linguistics and Literature, University of Bielefeld; 3IMS, University of Stuttgart

[moers, dmo@uni-bonn.de, petra.wagner@uni-bielefeld.de]

FAST SPEECH

Speech rate classification

1 2 3 4 5 6 7 8 9 18 19 20 21 22
slow normal fast ultra-fast

Fast speech characteristics:
Shortening, reduction, assimilation and elision of phones and syllables
- cause a worse perception of natural fast speech,
- are not applicable to produce fast speech in a unit selection synthesis system,
- therefore have to be avoided.

A speaker was selected
- who was a skilled speaker,
- who was able to speak both very fast and clear.

400 sentences were recorded in 2 conditions:
- normal,
- fast and clear.

EVALUATION

Linear acceleration by TD-PSOLA in 2 steps:
- normal rate sentences accelerated to tempo of fast rate sentences,
- normal as well as fast rate sentences accelerated to very fast rate (double fast rate).

Fast condition: normal rate stimuli judged to be more intelligible than natural fast spoken ones.

Very fast condition: fast rate stimuli clearly preferred with respect to naturalness.

DURATION PREDICTION (CART)

<table>
<thead>
<tr>
<th>Speech rate</th>
<th>Correlation</th>
<th>Mean (absolute Error (ms))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>0.82</td>
<td>20.16 (34.16)</td>
</tr>
<tr>
<td>Fast</td>
<td>0.78</td>
<td>12.37 (20.53)</td>
</tr>
</tbody>
</table>

No significant differences between normal and fast speech rate for the correlation between observed and predicated segment duration.

No significant differences in feature ranking between normal and fast speech.

REFERENCES


AUTOMATIC ALIGNMENT

49 sentences of each corpus labeled manually.
Label timing differences calculated by subtracting manual label time from automatic label time.

No significant differences between normal and fast speech rate for a 20 ms tolerance interval.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Correlation Normal</th>
<th>Correlation Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone identity</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>Phrase position</td>
<td>0.68</td>
<td>0.67</td>
</tr>
<tr>
<td>Follow. Phoneme</td>
<td>0.79</td>
<td>0.76</td>
</tr>
<tr>
<td>Preceding phoneme</td>
<td>0.80</td>
<td>0.77</td>
</tr>
<tr>
<td>Syllabic stress</td>
<td>0.81</td>
<td>0.77</td>
</tr>
<tr>
<td>2nd follow. phoneme</td>
<td>0.82</td>
<td>0.78</td>
</tr>
</tbody>
</table>