1. Overview

Personalized disfluent text generation
- Application: avatars speaking instead of humans
- Personalization: reproduce speakers’ individuality.
- Disfluency: reproduce human-like disfluency.
- Recent reading-style speech synthesis:
  - Synthesize only fluent speech.
- Personalized disfluent text generation (our goal)
  - Personalized disfluent text generation – this work
  - We focus on one kind of disfluency, filled pauses (FPs).
  - Ex. I’ll (uh) explain FP prediction.

Filled pauses (FPs): one kind of disfluency
- Roles: help speech generation [1] and communication [2].
- Diversity:
  - FP words (160 in Japanese [3])
  - Difference among speakers
- Features: position and word

This work
- Personalized FP generation by grouping speakers
- Improvement of prediction performance

2. Method

Basic architecture of FP prediction model [5]

Weighted cross entropy loss [6]
- Data imbalance:
  - It is harder to predict less frequent words.
- Increase weights of the loss of less frequent FP words.

Rich word representation model
- Use BERT [8] as rich word representation model.

Experimental conditions
- Criteria: precision / recall / F score / (specificity)
- Cross validation

Dataset
<table>
<thead>
<tr>
<th>137 speakers in CSJ [9]</th>
<th>LeCSPonSpeech [10]</th>
</tr>
</thead>
</table>
Tokenization
Word embedding
BERT (pretrained)
Prediction Model
BLSTM
Input Morphemes
Output 14 classes (none or 13 FP words)

Weighted cross entropy loss
- Using weighted loss improves the performance.

Rich word representation model
- BERT performs better than fastText.

Group-dependent models
- Higher scores than the universal model for both position and word, except for group 2 for position

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Universal</th>
<th>Grouping by word frequency</th>
<th>Grouping by position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
<td>0.376</td>
<td>0.454 0.456 0.427 0.390 0.461 0.323 0.413 0.444</td>
<td></td>
</tr>
<tr>
<td>Word</td>
<td>0.089</td>
<td>0.284 0.288 0.248 0.196 0.277 0.212 0.158 0.237</td>
<td></td>
</tr>
</tbody>
</table>

Speaker-dependent models
- Speaker models have lower scores.
- Speaker adaptation is difficult.
- Discussion
  - The universal model has higher scores than group-dependent models.
- For the prediction for such speakers, we can use the universal model.

References