LibriS2S: A German-English Speech-to-Speech Translation Corpus

Motivation and Background  Speech-to-Speech Translation (S2ST)
- Translate speech from one language to speech in another language.
- Current approaches (black/full arrows) use concatenation of Speech-to-Text (STT), text-to-text translation and Text-to-Speech (TTS) models.
- Can result in a loss of information on speech characteristics such as the pitch and energy (Sperber and Paulik, 2020).
- Proposed (blue/dashed arrows) pipeline to pass information from the source speech to the TTS system to synthesize the target speech:

```
Source Audio
<table>
<thead>
<tr>
<th>Additional information from source Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>STT</td>
</tr>
<tr>
<td>translation</td>
</tr>
<tr>
<td>TTS</td>
</tr>
</tbody>
</table>
```

- Requires parallel speech in both languages.
- Most datasets do not meet the requirement, require a costly license or have a limited amount of samples/no labeled data.
- This paper introduces LibriS2S for German and English, consisting of parallel speech and transcriptions.

Dataset Creation
Requirements: parallel human speech in both languages with their transcripts.
Steps:
- Start from existing dataset, LibrivoxDeEn (Beilharz et al., 2020), which contains the German speech, transcripts and English translation from audiobooks.
- Scrape the English audiobooks from librivox.org.
- Align the scraped data using speech alignment tools.
- Combine the scraped data and LibrivoxDeEn dataset.

<table>
<thead>
<tr>
<th></th>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td># Audio files</td>
<td>25 635</td>
<td>25 635</td>
</tr>
<tr>
<td># Unique tokens</td>
<td>10 367</td>
<td>9 322</td>
</tr>
<tr>
<td># Words</td>
<td>49 129</td>
<td>62 961</td>
</tr>
<tr>
<td># Speakers</td>
<td>42</td>
<td>29</td>
</tr>
<tr>
<td>Duration (hh:mm:ss)</td>
<td>52:30:57</td>
<td>57:20:10</td>
</tr>
</tbody>
</table>

Source Feature Vectors (SFV)
- Way to represent the pitch and energy (influences the volume and prosody of speech) for each phoneme.
- Used as additional input to give information on the speech characteristics in the source speech.
- Source sentence with pitch/energy values:

```
Source sentence with pitch/energy values:
```

```
Target sentence:
```

```
Target Phonemes with the mappd pitch:
```

- Pitch and energy mapped from word in source language to corresponding word/phonemes in target language if possible.

Model Adaptations of FastSpeech 2 TTS model
Selected FastSpeech 2 for its ability to control pitch and energy to a certain extent by the dedicated predictors.

```
Multiple adaptations of the FastSpeech 2 architecture have been tested:
- "phi": only uses the phoneme sequence from the source as additional input.
- "emb": only uses the SFV’s to replace the last 2 embedding dimensions.
- "epi": SFV’s used in the embedding and as input to the pitch and energy predictor.
- "addition": SFV’s added to the output of the energy/pitch predictor.
```

Models are trained on 2079 audio files from a single speaker and as vocoder an MB-MelGAN (Yang et al., 2020) model was finetuned on the same data.

Results
- MOSNet (Lo et al., 2019) used to approximate the Mean Opinion Score (MOS).
- Pitch moments and Dynamic Time Warping (DTW) to compare pitch to the Ground Truth (GT)
- Mean Absolute Error (MAE) for energy

<table>
<thead>
<tr>
<th></th>
<th>MOSNet</th>
<th>Pitch ( \sigma )</th>
<th>Pitch ( \gamma )</th>
<th>Pitch ( \kappa )</th>
<th>Pitch DTW</th>
<th>Energy MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GT</td>
<td>3.673</td>
<td>31.867</td>
<td>0.788</td>
<td>1.769</td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>3.133</td>
<td>41.163</td>
<td>-1.138</td>
<td>2.627</td>
<td>21.423</td>
<td>10.039</td>
</tr>
<tr>
<td>pho</td>
<td>3.161</td>
<td>40.778</td>
<td>-1.063</td>
<td>2.931</td>
<td>19.876</td>
<td>10.110</td>
</tr>
<tr>
<td>emb</td>
<td>3.163</td>
<td>38.113</td>
<td>-1.000</td>
<td>3.104</td>
<td>20.329</td>
<td>10.002</td>
</tr>
<tr>
<td>epi</td>
<td>3.159</td>
<td>38.704</td>
<td>-1.039</td>
<td>2.879</td>
<td>19.948</td>
<td>10.103</td>
</tr>
<tr>
<td>addition</td>
<td>3.071</td>
<td>42.174</td>
<td>-0.807</td>
<td>2.390</td>
<td>23.065</td>
<td>11.042</td>
</tr>
</tbody>
</table>

Conclusion
- New S2ST translation dataset released for the English, German language pair.
- The tools used are released together with it and can be utilized to extend it to other languages.
- Introduced dataset was used to train adaptations of FastSpeech 2 that also take information from the source speech as input.
- Results show that the adapted models improve upon the baseline model but need further investigation.

References