1. Background

**Code-switching (CS)**: Bilingual speakers switch languages within a conversation.

**CS challenges**: Handle the input in a multilingual setting.

**Existing approaches**:
- Train either only ASR or only TTS.
- By supervised learning with CS data.

**Problems**: Parallel speech & text CS data are generally unavailable.

**Speech chain** [Andros et al., 2017]

- **Human speech chain**: A closed-loop mechanism has a critical auditory feedback mechanism.
- **Machine speech chain**: A closed-loop architecture which allows ASR and TTS to teach each other.

2. Proposed Method

**Based on how humans learn CS**:
- Listening while speaking CS using a speech chain framework.
- Enable to perform semi-supervised learning.
- Aim to improve both ASR and TTS at the same time.

**Code-switching with Speech Chain**

**Step 1. Supervised learning**
Separately train ASR & TTS with parallel speech-text monolingual data.

**Step 2. Unsupervised learning**
Perform a speech chain with only CS text or CS speech.

**Training objective**: Possible to train the new matters without forgetting the old one.

3. Experimental Results

**Set-up**:
- **Model**: Attention-based encoder-decoder ASR, Tacotron TTS.
- **Data**: En/Ja monolingual BTEC text data.
- En-Ja CS text created from BTEC, Speech is synthesized by GoogleTTS.

**ASR & TTS Performances**
(in CER & L2-norm squared in log-Mel spectrogram)

<table>
<thead>
<tr>
<th>Model</th>
<th>Japanese test(JaTTS)</th>
<th>Japanese test(MixTTS)</th>
<th>English test(EnTTS)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ASR</td>
<td>TTS</td>
<td>ASR</td>
</tr>
<tr>
<td>Baseline: paired speech-text ⇒ Supervised training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ja50k(JaTTS)</td>
<td>2.11%</td>
<td>0.321</td>
<td>33.73%</td>
</tr>
<tr>
<td>En50k(EnTTS)</td>
<td>86.42%</td>
<td>0.373</td>
<td>66.16%</td>
</tr>
<tr>
<td>Ja25k + En25k(MixTTS)</td>
<td>1.71%</td>
<td>0.312</td>
<td>18.11%</td>
</tr>
<tr>
<td>Speech chain: [paired Ja25k+En25k (MixTTS)] + [unpaired CS (Mix+JaTTS)] ⇒ Semi-supervised training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+CS20k(Ja+MixTTS)</td>
<td>1.82%</td>
<td>0.305</td>
<td>5.08%</td>
</tr>
</tbody>
</table>

**Note**: MixTTS means using both JaTTS and EnTTS.

4. Conclusion

**Proposed CS Model** based on speech chain:
- Allows CS ASR & CS TTS to learn from each other.
- Even without any parallel speech & text CS data.

**Experimental results reveal**:
- Maintaining performance in the monolingual setting.
- Improved ASR in CS test from CER 18.11% to 5.08%.
- Also improved TTS from L2-norm 0.489 to 0.372.

**Future Work**:
- Use natural speech data.
- Apply to other languages.