1. Introduction

- Speech chain model integrates ASR and TTS into a single cycle during training.
- By combining both models, we could train with auxiliary feedback loss.
- Problem:
  - The output from ASR are discrete tokens
  - Non-differentiable (ASR → TTS)
- Solution:
  - Apply straight-through estimator on Gumbel-softmax or argmax sample

2. Speech Chain and Feedback loss

```
x = \dot{y} \quad \Rightarrow \quad x = \hat{y} \quad \text{ST}
```

Feedback loss: \( L_{TTS} = || x - \hat{x} ||^2 \) where \( x = TTS(\hat{y}, z) \)

- a) Speech chain loop with speaker embedding module.
- b) Original: feedback \( L_{TTS} \) can’t be backpropagated through variable \( \hat{y} \).
- c) Proposal: Estimate gradient through variable \( \hat{y} \) with straight-through estimator.

3. Straight-through Estimator (ST)

- ST-argmax
  - Deterministic choosing token by highest probability.
  - \( \hat{y}_t = argmax_c \ p_{y_t}[c] \)

- ST-Gumbel softmax
  - Sampling a token from \( p_{y_t}[c] \):
    \[
P_{y_t}[c] = \frac{\exp\left(h_t^d[c]/\tau\right)}{\sum_{c=1}^C \exp\left(h_t^d[c]/\tau\right)}
\]
  - \( \tau = \text{temperature} \)

4. Experiment

<table>
<thead>
<tr>
<th>Baseline Model</th>
<th>CER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att MLP</td>
<td>7.12</td>
</tr>
<tr>
<td>Att MLP-MA</td>
<td>6.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Proposed Model</th>
<th>Generation</th>
<th>ST</th>
<th>CER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Att MLP-MA</td>
<td>Teacher-forcing</td>
<td>argmax</td>
<td>5.75</td>
</tr>
<tr>
<td>Att MLP-MA</td>
<td>Teacher-forcing</td>
<td>gumbel</td>
<td>5.7</td>
</tr>
<tr>
<td>Att MLP-MA</td>
<td>Greedy</td>
<td>argmax</td>
<td>5.84</td>
</tr>
<tr>
<td>Att MLP-MA</td>
<td>Greedy</td>
<td>gumbel</td>
<td>5.88</td>
</tr>
</tbody>
</table>

5. Discussion

- We introduced ST-estimator for training ASR module based on TTS feedback loss.
- The gradient in discretization problem can be replaced by identity Jacobian matrix.
- Our experiment shows that by adding auxiliary feedback loss, we improve the ASR performance further.