**Introduction**

- Aim to ease a Bible study group everyday life.
- Transcribe Bible teachings in audio form recorded by Pastor Wang in Chinese mandarin to searchable text.

Unique in several ways:
- Focus on biblical context in mandarin
- Limit scope to a single speaker, Pastor Wang

Generalization challenges:
- Overlooked accented mandarin speaking
- Missed Bible terminology
- Went astray from coherency of Bible teachings

Take on the challenges:
- Leverage end-to-end deep learning model, DeepSpeech2, to investigate Automatic Speech Recognition (ASR) system
- Use Single Character Error Rate (CER) metric
- Compare different model architecture
- Apply larger dataset for model training
- Tune hyperparameter, alpha and beta
- Accelerate model training time

Project Goal: Develop a practical ASR system which can transcribe Pastor Wang’s Bible teachings in Chinese mandarin. The CER is aimed for 5% and below.

Results: We successfully structured and applied strategy of deep learning to drop CER from 50.39% to 24.97%. Although the final CER does not meet initial goal of 5%, we have identified several areas for future to further improve the CER.

**Dataset and Feature**

- Two dataset (aishell) and (aishell2) are used for model training.
- Pastor Wang’s teachings are used as Test Set.
- Dataset is divided according to its size.

<table>
<thead>
<tr>
<th>Dataset Size (utterance)</th>
<th>Dataset (aishell)</th>
<th>Dataset (aishell2)</th>
<th>Pastor Wang (Teachings)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset Size (hour)</td>
<td>141,925</td>
<td>1,000,922</td>
<td>226</td>
</tr>
<tr>
<td>Training Set (%)</td>
<td>84.8</td>
<td>99.0</td>
<td>0</td>
</tr>
<tr>
<td>Training-dev Set (%)</td>
<td>10.1</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>Training-test Set (%)</td>
<td>5.1</td>
<td>0.5</td>
<td>100</td>
</tr>
<tr>
<td>Test Set (%)</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
</tbody>
</table>

Illustration of an audio clip in format of waveform
- Encoding eight Chinese mandarin characters

- DeepSpeech2 may include Conv layers from 1 to 3, and GRU from 1 to 7
- We explored 2-Conv, 3-GRU layers and the maximum capacity 3-Conv, 7-GRU
- Using multiple GPUs is not accelerating training in our use case
- To avoid running out GPU memory
  - GRU size uses 1024
  - Mini-batch size is 16

**Experiments Results Discussion**

- Poor initial CER on DeepSpeech2
- State-of-the-art benchmark CER on AWS

Error Analysis:
- Output length is not matched with target
- Transcribed character is not accurate

Target Transcription: 讲到神伟大的主权
Output Transcription: 将到人轨道的主持
Character Error Rate [CER]: 62.50%

CER Improvement:
- Larger dataset (aishell2) for model training
- Hyperparameter (alpha, beta) tuning
  - Dropped more than 20% CER
- Larger network helps not much

**Conclusion**

- An elaborate engineering task to apply deep learning
- Data preprocessing is mandatory for audio to fit model
- Large training dataset matters the most
- Tuning hyperparameter and network architecture matter
- Future work to train efficiently and further lower CER