Code for the paper: "Oracle Teacher: Towards Better Knowledge Distillation" Speech Recognition Task

1. Directory Structure

```
Oracle_Teacher_Training
Student_KD_Training
Student_CTC_Training
```

2. Get started

This work was tested with Tensorflow (Tensorflow-gpu) 1.14.0, CUDA 10.1, python 3.7, and Ubuntu 16.04. Other necessary packages listed in requirements.txt. Installation instructions regarding OpenSeq2Seq toolkit are provided by https://nvidia.github.io/OpenSeq2Seq/html/index.html

3. Training the Oracle Teacher

```
cd Oracle_Teacher_Training
```

1) Download and preprocess LibriSpeech dataset

```
sudo apt-get -y install sox libsox-dev
mkdir -p data
python scripts/import_librivox.py data/
    librispeech
```

2) Train the Oracle Teacher

```
CUDA_VISIBLE_DEVICES=0 python run.py --
    config_file=example_configs/
    speech2text/oracle_teacher.py --mode=
    train
```

3) Extract the knowledge from the Oracle Teacher

```
CUDA_VISIBLE_DEVICES=0 python run.py --
    config_file=example_configs/
    speech2text/oracle_teacher.py --mode=
    infer
```

You can change the saving path in the following file:

```
./open_seq2seq/models/speech2text.py
```

4. Initializing the Student with KD

```
cd Student_KD_Training
```

1) Preprocess the data

```
mkdir -p data/librispeech
bash preprocess_kd_csv.sh ../
    Oracle_Teacher_Training/data/
    librispeech/librivox-train-clean-100.
    csv ./data/librispeech/librivox-train
    -clean-100.csv
bash preprocess_kd_csv.sh ../
    Oracle_Teacher_Training/data/
    librispeech/librivox-train-clean-360.
    csv ./data/librispeech/librivox-train
    -clean-360.csv
bash preprocess_kd_csv.sh ../
    Oracle_Teacher_Training/data/
    librispeech/librivox-train-other-500.
    csv ./data/librispeech/librivox-train
    -other-500.csv
```

2) Train the Student Model w/ FitNets

```
CUDA_VISIBLE_DEVICES=0,1,2 python run.py
     --config_file=example_configs/
    speech2text/jasper_mini_3gpu.py --
    mode=train
```

5. Training the Student with CTC

```
cd Student_CTC_Training
```

1) Move the files

```
mv ../Student_KD_Training/Student_KD ./
cp -r ../Student_KD_Training/data ./data
```

2) Train the Student Model w/ CTC

```
CUDA_VISIBLE_DEVICES=0,1,2 python run.py
     --config_file=example_configs/
    speech2text/jasper_mini_3gpu.py --
    mode=train_eval
```

3) Test the Student Model

```
CUDA_VISIBLE_DEVICES=0,1,2 python run.py
     --config_file=example_configs/
    speech2text/jasper_mini_3gpu.py --
    mode=eval
```