Outline

- Task Description
- Dataset
- Data segmentation
- Hints
- Kaggle

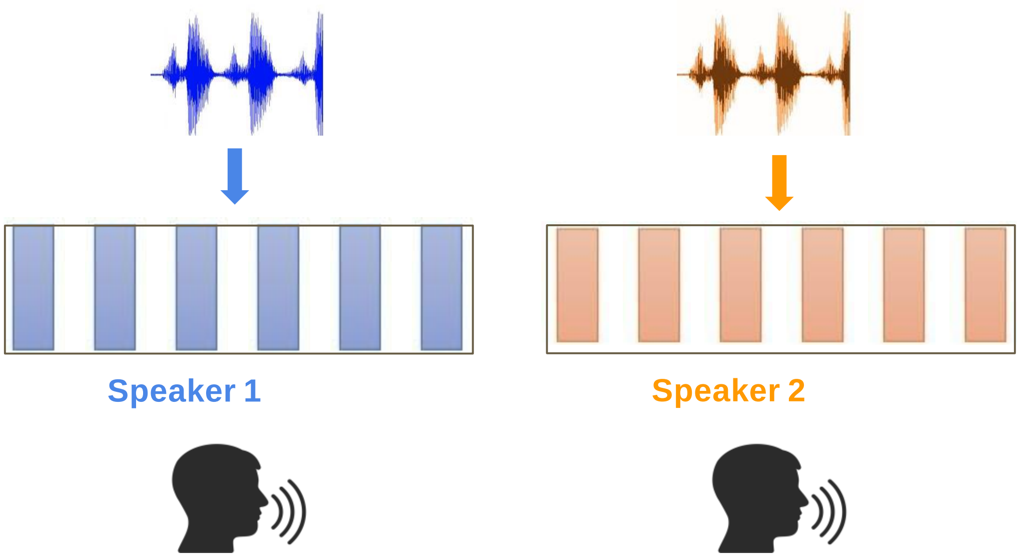
Task Introduction

- Self-attention
  - Proposed in GOOGLE's work, Attention is all you need . It combines the strengths of RNN (consider whole sequence) and CNN (processing parallelly).
- Main goal: Learn how to use transformer.

HW4: Speaker classification

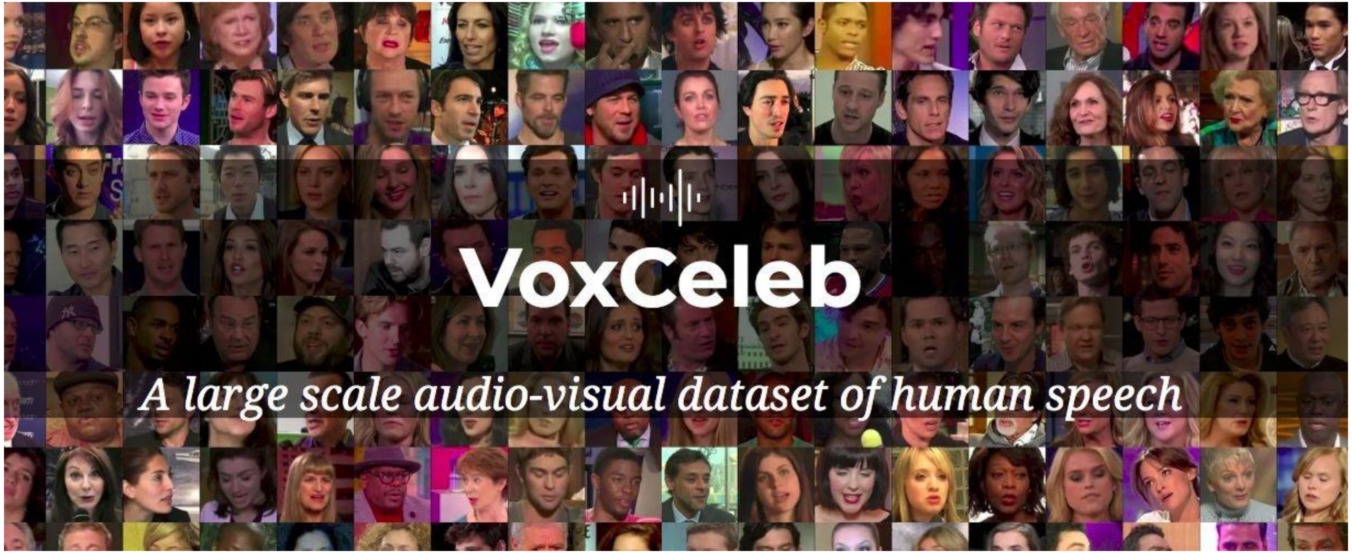
Task: Multiclass Classification

Predict speaker class from given speech.

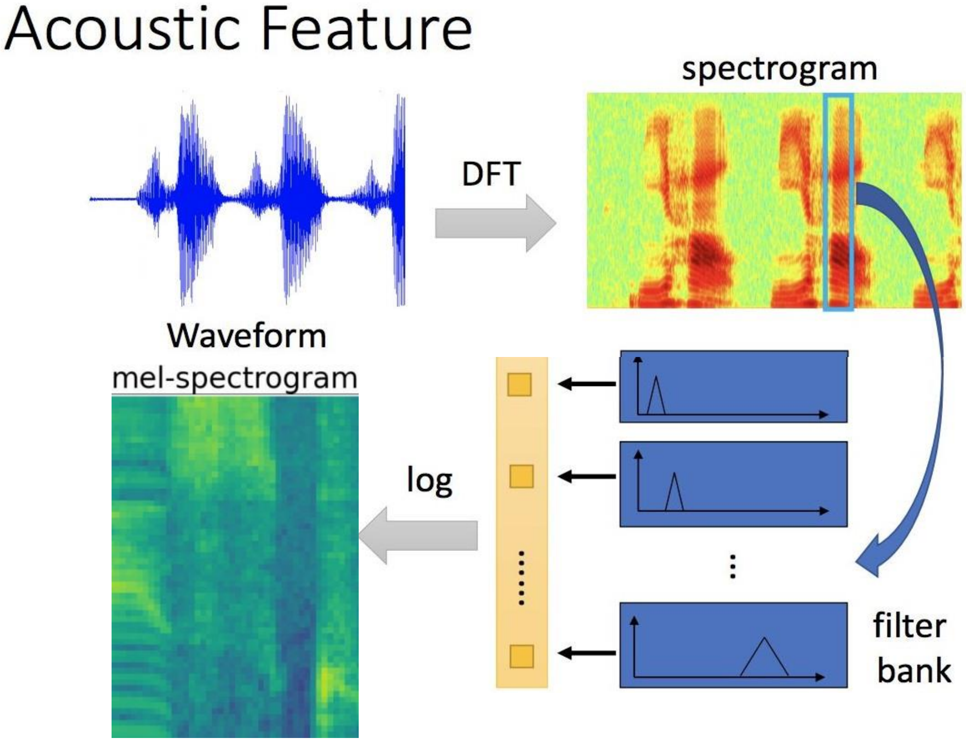


Dataset

- Training: 62783 processed audio features with labels.
- Testing: 6656 processed audio features without labels.
- Label: 600 classes in total, each class represents a speaker.



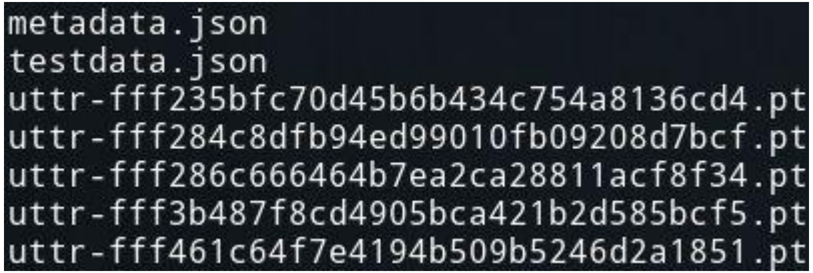
Data Preprocessing



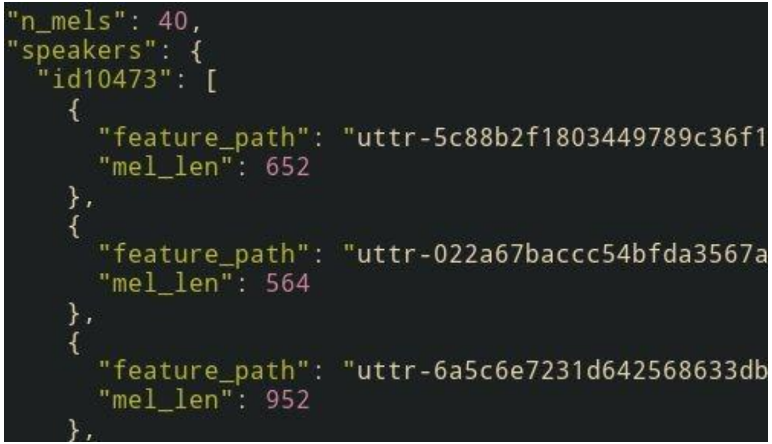
ref.
prof. Hung-Yi Lee
[2020Spring DLHLP] Speech Recognition

Data formats

- Data Directory
  - metadata.json
  - testdata.json
  - mapping.json
  - uttr-{random string}.pt
- The information in metadata
  - "n_mels": The dimention of mel-spectrogram.
  - "speakers": A dictionary.
    - Key: speaker ids.
    - value: "feature_path" and "mel_len"

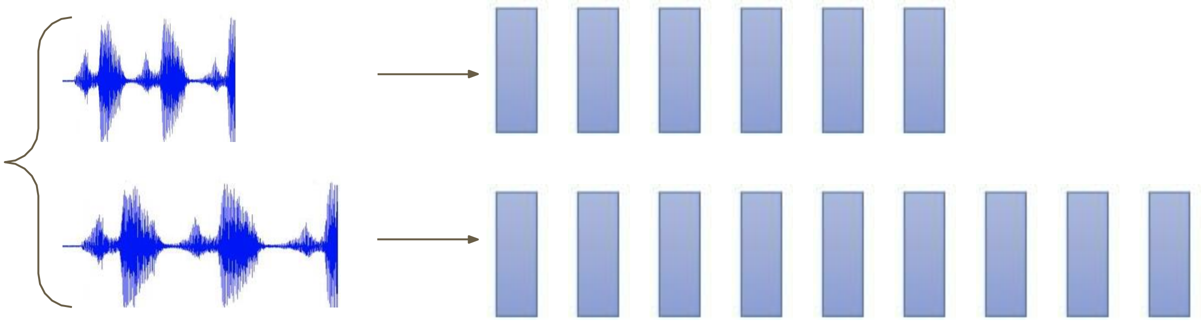


```
metadata.json
testdata.json
uttr-fff235bfc70d45b6b434c754a8136cd4.pt
uttr-fff284c8dfb94ed99010fb09208d7bcf.pt
uttr-fff286c666464b7ea2ca28811acf8f34.pt
uttr-fff3b487f8cd4905bca421b2d585bcf5.pt
uttr-fff461c64f7e4194b509b5246d2a1851.pt
```



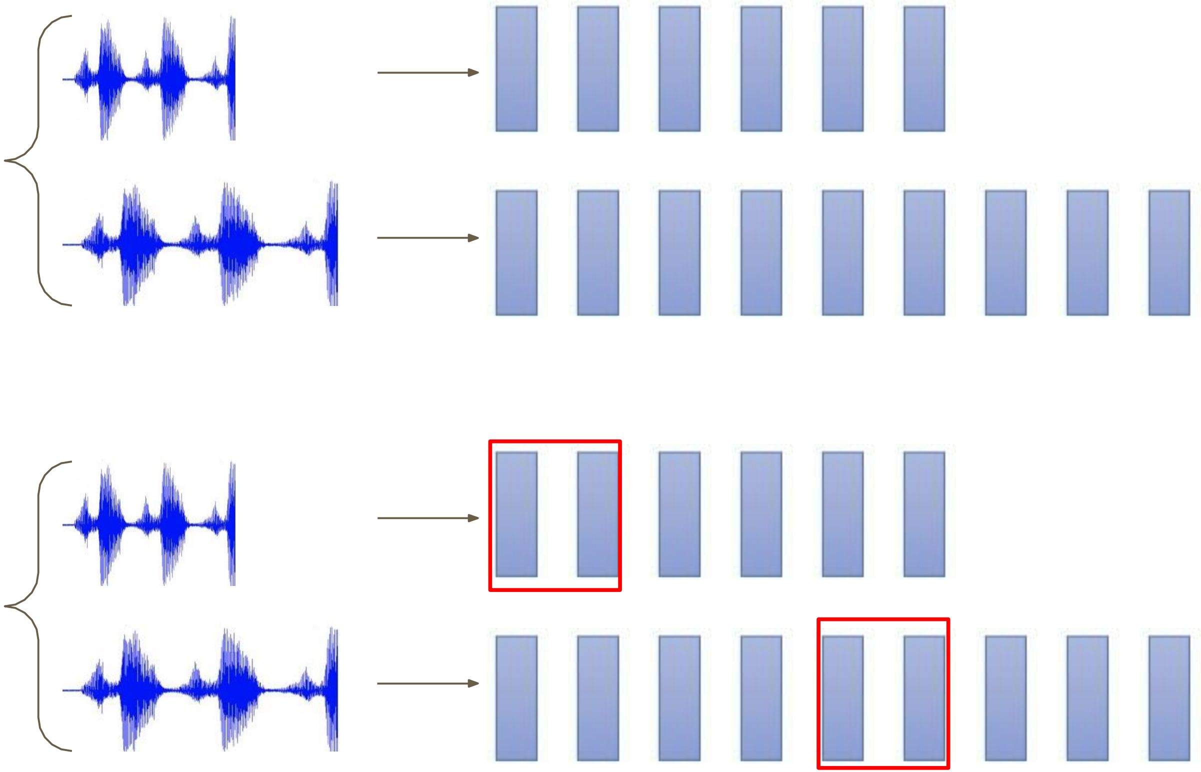
```
"n_mels": 40,
"speakers": {
  "id10473": [
    {
      "feature_path": "uttr-5c88b2f1803449789c36f1
      "mel_len": 652
    },
    {
      "feature_path": "uttr-022a67baccc54bfda3567a
      "mel_len": 564
    },
    {
      "feature_path": "uttr-6a5c6e7231d642568633db
      "mel_len": 952
    },
```

Data segmentation during training



Different length

Data segmentation during training



Different length

Segment during training

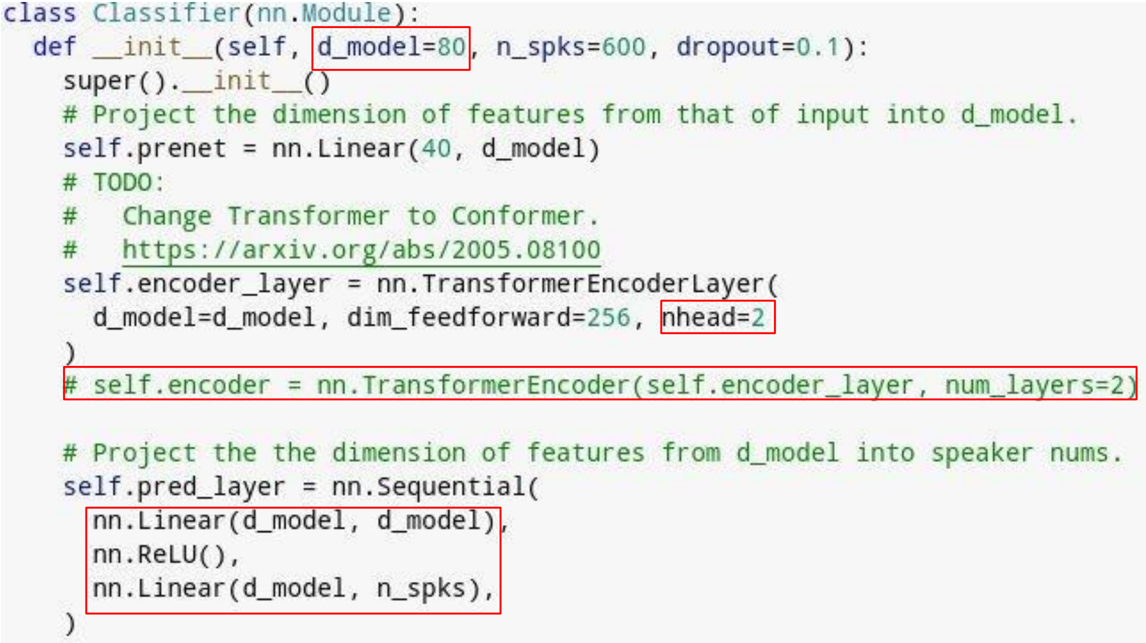
Segment = 2

Hints

- Simple: Run sample code and know how to use transformer.
- Medium: Know how to adjust parameters of transformer.
- Hard: Construct conformer which is a variety of transformer.

Hints

- Modify the parameters of the transformer modules in the sample code.

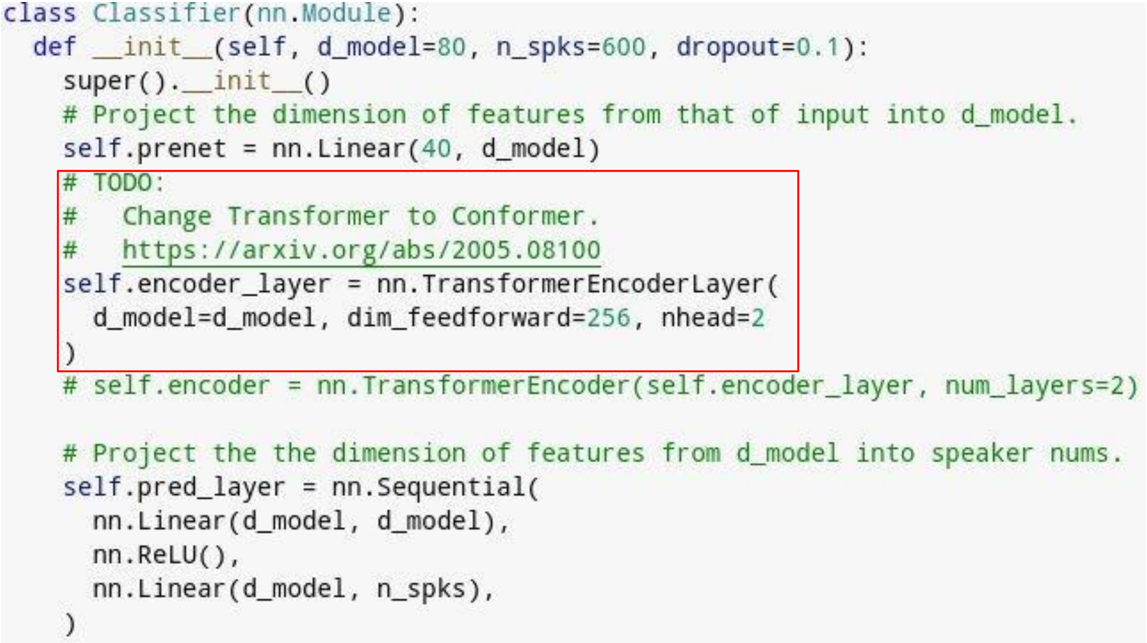


```
class Classifier(nn.Module):
  def __init__(self, d_model=80, n_spks=600, dropout=0.1):
    super().__init__()
    # Project the dimension of features from that of input into d_model.
    self.prenet = nn.Linear(40, d_model)
    # TODO:
    #   Change Transformer to Conformer.
    #   https://arxiv.org/abs/2005.08100
    self.encoder_layer = nn.TransformerEncoderLayer(
      d_model=d_model, dim_feedforward=256, nhead=2
    )
    # self.encoder = nn.TransformerEncoder(self.encoder_layer, num_layers=2)

    # Project the the dimension of features from d_model into speaker nums.
    self.pred_layer = nn.Sequential(
      nn.Linear(d_model, d_model),
      nn.ReLU(),
      nn.Linear(d_model, n_spks),
    )
```

Hints

- Improve the performance by constructing the conformer layer.

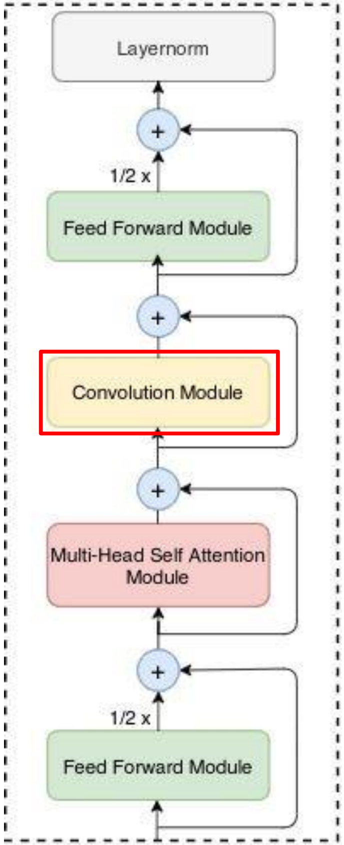


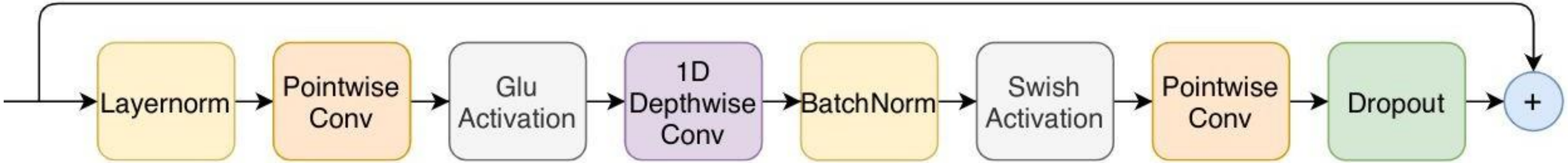
```
class Classifier(nn.Module):
  def __init__(self, d_model=80, n_spks=600, dropout=0.1):
    super().__init__()
    # Project the dimension of features from that of input into d_model.
    self.prenet = nn.Linear(40, d_model)
    # TODO:
    #   Change Transformer to Conformer.
    #   https://arxiv.org/abs/2005.08100
    self.encoder_layer = nn.TransformerEncoderLayer(
      d_model=d_model, dim_feedforward=256, nhead=2
    )
    # self.encoder = nn.TransformerEncoder(self.encoder_layer, num_layers=2)

    # Project the the dimension of features from d_model into speaker nums.
    self.pred_layer = nn.Sequential(
      nn.Linear(d_model, d_model),
      nn.ReLU(),
      nn.Linear(d_model, n_spks),
    )
```

Hints

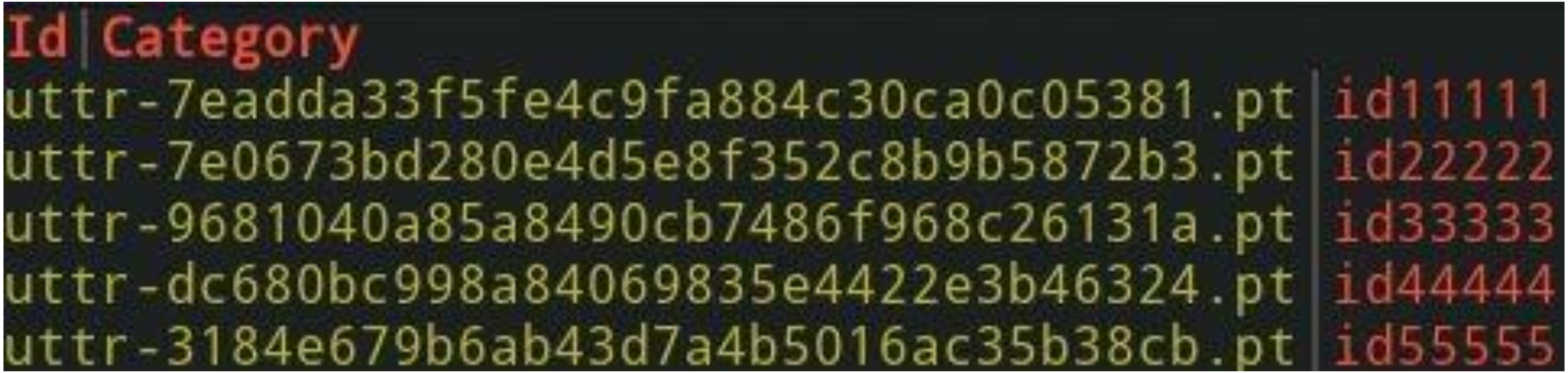
Conformer：https://arxiv.org/abs/2005.08100





Submission Format

- "Id, Category" split by ',' in the first row。
- Followed by 6666 lines of "filename, speaker name" split by ','.



```
Id|Category
uttr-7eadda33f5fe4c9fa884c30ca0c05381.pt|id11111
uttr-7e0673bd280e4d5e8f352c8b9b5872b3.pt|id22222
uttr-9681040a85a8490cb7486f968c26131a.pt|id33333
uttr-dc680bc998a84069835e4422e3b46324.pt|id44444
uttr-3184e679b6ab43d7a4b5016ac35b38cb.pt|id55555
```