**课程大作业要求**

**1. 任务： Automated Speech Recognition**

语音识别（Automatic Speech Recognition，ASR）旨在将人类的口头语言转换为文本形式。它是自然语言处理（NLP）领域中的一个重要研究方向。ASR系统接受来自语音输入源（如麦克风、电话或语音文件）的声音信号，并将其转换为相应的文本输出。

参考文献：

Malik, Mishaim, et al. "Automatic speech recognition: a survey." Multimedia Tools and Applications 80 (2021): 9411-9457.

**2. 数据集**

LibriSpeech 100 hours

下载链接：<https://www.openslr.org/12>

Train: <https://www.openslr.org/resources/12/train-clean-100.tar.gz>

Val: <https://www.openslr.org/resources/12/dev-clean.tar.gz>

Test: <https://www.openslr.org/resources/12/test-clean.tar.gz>

注: LibriSpeech是从电子书网站LibriVox采集的，总共包含1000小时的语音数据。但本作业出于训练难度考虑，只需要用其中的100小时。

**3. 参考方法以及基础知识的补充**

**3.1 基础知识**

首先我们按照Task formulation简单介绍完成本次作业需要的基础知识。但建议大家阅读Sec 1中的survey，以更加全面深入了解ASR的发展。

语音信号通常是按照一定的采样率保存为数字信号。本作业中，我们统一采样率为16K。我们可以通过以下的方法load语音：

|  |
| --- |
| import librosa  wav\_path = “example.wav” # or mp3 … formats  sample\_rate = 16000 # 16K sample rate  wav\_data, \_ = librosa.load(wav\_path, sr=sample\_rate) |

作为ASR模型的输入，我们通常需要将raw data (wav\_data)转换成特殊的特征，我们在本作业中使用MFCC格式[1], 可以通过以下的代码提取特征：

|  |
| --- |
| mfcc\_feats = librosa.feature.mfcc(y=wav\_data, n\_mfcc=80, sr=16000)# n\_mfcc denotes the dimension of mfcc feature. Usually n\_mfcc=20, 39, 80. We recommend using 80 for better results (20 is OK if you are limited by resources). |

ASR模型的输出通常是自然语言(a sequence of tokens)，可以参考机器翻译 (Neural Machine Translation)。因为本作业是英文的数据集，BPE是一个推荐选择 (SentencePiece package is recommended.)。

ASR的evaluation metric是word error rate (WER), 计算方法详见survey paper的Sec 3.2.3。本课程已经为大家提供了统一了评测工具：<https://github.com/AlanSwift/asr_course_tools>，请按照github的readme对你实现的ASR系统进行性能评测。

**3.2 推荐方法**

通常ASR模型是encoder decoder架构的。其中decoder按照不同的方法，通常可以分成hybrid model，end-to-end model。按照这个分类，我们给出以下的推荐模型：

**3.2.1 Encoder**

a) Transformer:

参考论文：Attention Is All You Need

Transformer-based Acoustic Modeling for Hybrid Speech Recognition

b) Conformer:

参考论文：Conformer: Convolution-augmented Transformer for Speech Recognition

通常来说，transformer比较省资源，conformer更加耗GPU显存，但能取得更好的效果。

**3.2.2 Decoder:**

**1） End-to-End Diagram**

a) Transformer decoder

参考论文：Attention Is All You Need

普通的transformer decoder，可以仿照自回归机器翻译的decoder做法。

b) Transducer

transducer是一种高效的ASR decoder，具有流式输出能力。 通常可以用RNN-transducer和transformer-transducer。如果你选用transducer，这两种都是可以的。

参考论文：Transformer Transducer: A Streamable Speech Recognition Model with Transformer Encoders and RNN-T Loss

**2）Hybrid Diagram**

Hybrid diagram通常指的是在attention based decoder (i.e., transformer decoder in 4.2.2’s part 1)的基础上，加入一个language model，以生成更高质量的ASR transcription。Language model可以是n-gram，transformer-LM等选择。

参考文献：

Connectionist speech recognition: a hybrid approach

Transformer-based Acoustic Modeling for Hybrid Speech Recognition (Sec. 2, Hybrid Architecture)

本次作业只需要选择一种组合方式实现即可。

References

1) <https://en.wikipedia.org/wiki/Mel-frequency_cepstrum>

Contact me ([shenkai@zju.edu.cn](mailto:shenkai@zju.edu.cn)) if you have any questions.