# 陈培杰

求职意向：
语音处理，自然语言处理

**邮箱** peijie_chen_sz@163.com
**地址** 中国，广东省，深圳市
**电话** 7422583873(英) 15521029020(中)
**博客** https://markchanxdjent.github.io/


## 技能

**熟悉语音处理以及NLP的相关算法** **Python 编程经验**

**熟悉Linux的操作环境** **熟悉Kaldi和Pytorch等相关工具**

## 语言

| | |
|---|---|
| 中文 | 母语 |
| English | Advanced |

## 教育

**爱丁堡大学**
**语音与语言处理 硕士**
英国，爱丁堡
2018-2019

1. 修习语音处理，自然语言处理，机器学习等相关课程
2. 第一学期成绩良好(Merit)
3.毕业设计：Long-term sequence models for lightly-supervised data ；该毕业设计与爱丁堡的语音科技公司Quorate 合作，主要研究LSTMs等long-term sequences model在半监督学习上的应用

**华南师范大学**
**语言学 学士**
中国，广州
2014-2018

1. GPA 3.8/5.0
2. 本科期间曾发表语言学相关的论文
3. 本科论文通过爬取网易云音乐上的歌词数据探究华语歌曲中的语码混用现象

## 项目

***基于Kaldi的英语语音识别系统***

1. 通过Kaldi建立GMM-HMM以及DNN-HMM语音识别系统，训练和测试的数据来源为TIMIT，同时，也用自己录制的文件进行测试。
2. 对比不同参数，例如不同的高斯模型以及聚类数量对WER的影响。
3. 通过'gender_map'建立一个gender-dependent的模型，既提高系统对某一性别的友好度。
4. 对比GMM-HMM模型和DNN-HMM模型在说话人自适应任务上的表现有何差异。探究i-vector的引入能多大程度提升系统在说话人自适应上的表现。

***基于HTK的英语数字语音识别器***

1. 用于识别数字0-9的GMM+HMM语音识别器。
2. 使用HTK工具搭建，训练和测试用的数据皆由自己以及同学录制完成
3. 通过实验研究影响识别WER的相关因素，包括麦克风的种类，口音，性别以及HMMs中的state的数目

***Seq2Seq 日语-英语神经网络翻译器***

1. 该项目是对一个现有的Seq2Seq翻译器进行改进。原有的翻译器是一个在编码器和解码器上分别使用双向lstms以及单向lstms的Seq2Seq模型，该项目通过引入lexical model和注意力机制来提升翻译器的表现。
2. 注意力机制和lexical model的框架来源于两篇论文，论文复现之后，通过实验研究超参对翻译结果的影响。
3. 同时，该项目也发现了lexical model对于低资源语言的翻译有一定的提升，因为lexical model在处理unknown words的表现更好。

***Uni selection 语音合成（基于Festival）***

1. Uni selection合成可以简单理解为给每个diphone添加不同的候选，而diphone合成里的每个diphone只有一个候选。
2. Festival是爱丁堡大学开发的语音合成系统，可以用于载入此项目录制的数据。
3. 在基本的语音合成系统搭建成功之后，通过相关的实验来探究影响合成效果的因素，例如data base的大小，diphone的覆盖度，调整target cost和join cost的权重等。

***Diphone 语音合成***

1. 利用python搭建的英语语音合成系统。
2. Front End使用NLTK工具处理用户输入的文本。主要处理一些Non-standard Word，例如将“Dr.“扩展为“Doctor”等。之后再将输入的问题本拆分成diphone的合集，并且传入到Back End进行合成。
3. Back End根据Front End输出的diphone合集挑选对应的diphone语音文件，将其转换为numpy格式之后再把diphone拼接起来，最后再出成waveform文件播放。

***通过RNN预测动词单复数***

1. 该项目首先研究如何将反向传播算法以及随时间反向传播算法（BPTT）引入到RNN中。之后再将RNN用于词预测，该项目探究不同算法对词预测准确度的影响。
2. RNN同时也可以用来预测动词的单复数，该项目通过相关实验，探究动词和名词之间的attractors数量对预测准确度的影响。

***Twitter数据研究***

1. 学校提供Twitter的相关数据
2. 通过不同的实验探究不同相似度计算方法的优缺点。
3. 探究词频对不同的相似度计算方法的影响。

***爬取网易云音乐歌词***

1. 该项目主要通过python的beautiful soup4 以及github上一些网友分享的网易云apk爬取了部分华语歌手的歌词（总计7900首）。
2. 之后再对这些歌词进行语言学研究，分析他们歌曲中的语码混用及其特点和原因。