实习报告

**标题：**语音与多模态实验室实习

**摘要：** 语音识别是当前一个比较热门的领域，任务是将语音信息通过语音识别的模型解码成文本信息。语音关键词检索是语音检索的一个子方向。从2017年7月7日开始至今，本人在电信院李明老师的实验室中学习语音识别和关键词检索，同时做一些工程，训练一些模型。期间主要使用Kaldi进行模型训练和关键词检索。

**关键词：**语音识别 Kaldi 深度学习 WFST 语音关键词检索

**正文**

一、实践工作概况

在实习过程中的主要任务如下所述：

1. 使用Kaldi[1]进行语音识别系统的构建。Kaldi是语音识别当中非常常用的库，可以用于进行大规模连续语音识别模型的构建。语音识别系统由以下几个部分构成[2]：声学模型，语言模型，解码器。声学模型主要是计算对应从语音信号到音素的概率，而语言模型则是计算从音素信息到到文本的概率。解码器将前二者结合起来，构建HCLG[3]网络图。进行语音识别解码的时候就是在图中使用维特比[4]算法得出最近的一批结果，然后再在结果中挑选最合适的结果输出到文本。

构建声学模型首先需要对语音信号进行特征提取。一般而言，选择MFCC[5]作为语音的特征能够达到一个比较好的效果。构建声学模型的时候一般是针对基于上下文关联的音素进行建模，然后使用HMM[6]把不同的模型通过转移概率联系起来。训练声学模型，我们可以选择训练多高斯模型的参数，也可以训练选择深度神经网络的权重。后者的词错误率会更加低。

语言模型的构建我们使用n元语法[7]，即第n个出现的词只与前面n-1个词相关。一般而言，我们使用三元或者四元就能达到一个比较好的效果。

解码器的构建需要使用到WFST（加权有限状态转移）。它可以将语音识别的各个环节的模型结合起来，综合每一个部分的得分，得到最好的输出路径。解码器组成是HCLG，分别是HMM模型，上下文关联的音素，单词的音素组成，语言模型。使用者讲语音信号输入，就能都得到系统综合得分最高的结果。

2. 针对中国人说英语的模型

我们训练了针对中国人说英语的模型。数据集组成如下：115G语音数据作为训练集，5G语音数据作为测试集。参考Kaldi中实现的Librispeech的方法，我们自己写了脚本。Kaldi构建系统的流程如下：首先，对语音数据进行数据准备。训练一个好的模型需要大量数据，把数据全部放入内存是不现实的，所以我们需要构建一些索引，方便我们查找到语音数据以及其特征，这就是数据准备的一个好处。除此之外，还需要给声学模型准备单词列表和它们的音素组成，还需要给语言模型准备一些文本等等。数据准备完毕之后，就可以开始提取语音信号的MFCC特征，提取完特征之后再进行归一化处理，得到我们声学模型训练的材料。训练的过程是这样的，我们先使用传统的方法训练数据，得到一个初始的划分，最后再在当前语音划分的基础上进行进一步的训练。数据的添加需要慢慢进行，然后根据每一次训练的模型结果进行细节的调整。语言模型同样地，也需要训练。最后，利用生成的模型构建HCLG解码网络，然后再进行解码。后来我们又添加了新的Librispeech数据和Fisher数据，训练了一个更大的模型。

3. 关键词检索系统的构建

语音关键词检索的含义是，从一个语料库中寻找到我需要的词的位置的方法。给定一个词（以文字的形式），或者给一个语音片断之后，系统就找出该信息在语料库当中出现的位置。目前主流的方法是在大规模词表连续语音识别系统[8]的基础上构建的，然后构建WFST索引，最后找出关键词的位置。大规模词表连续语音识别系统在识别语音的时候会生成网格[9]，然后在网格中寻找最优路径，这个最优路径就是语音识别的结果。但是，这个识别出来的结果只是一个最优的结果，有时候关键词可能就潜藏在其他可能的路径上。为了把这些潜藏的结果也找出来，我们就要在网格上进行检索。关键词检索是序列中寻找词的过程。序列到结果的问题可以引入WFST[10]。我们可以构建一个索引，将结果的后验概率值，词的时间起点，词的时间终点，三个信息放在权重上。这个转移上的符号称之为时间因子。最后，我们将要检索的关键词合并到索引当中然后进行优化，排序，就能够找到关键词的相应位置和权重值。

Kaldi中关于关键词检索的实现基本原理就是上面的内容，在这里，我根据Kaldi中的Babel系统，写了脚本进行检索。

完成这些任务的意义要从两个方面来阐述，从实际应用的角度，我们的模型可以应用在英语教育之上，而更大的意义在于自身学习，自我提高上。完成这些任务对于今后个人的发展也会有好处。

二、工作方法与实施步骤

在实践中我的实施步骤如下：

1. 通过课件和网络资料进行基础知识的学习。

熟练使用Kaldi需要对Kaldi的搭建方法有足够的了解，这要求使用者掌握基本的语音识别知识和Linux脚本的编写能力。语音识别的基础知识，我重点看了老师的课件以及网上相关的博客进行学习。Linux主要看awk，sed，perl相关的知识点。

1. 进行实际操作，训练语音识别模型。

语音的相关实践主要集中在对Kaldi的使用上。学习Kaldi可以先从最基础的tidigits系统开始，先了解语音识别基础系统的结构。然后，通过wsj和librispeech系统，了解如何用神经网络训练大量语音数据。深入学习了这几个系统之后，我尝试跑通hkust系统。

之后，就是进行正式的模型训练了。我们编写脚本，对中国人说英语的数据进行了处理及训练。后面，我们又添加入了Librispeech和fisher的数据。

1. 深入关键词检索，开始看一些论文。

从十一月份开始，我转向语音关键词检索的学习。我通过论文了解该领域的常用方法，学习相关的基本步骤。了解了基本方法之后，我开始学习Kaldi中的Babel脚本，进行关键词检索系统的搭建。我主要在两个系统上进行关键词检索。一个系统是使用Cantonese数据训练的系统，另一个系统使用的是CVTE的开源模型。

工作效果

1. **语音识别模型**

我在实验室实践期间和小组的其他成员一起训练了一些模型，下面对模型的效果进行一个简单的总结和统计。如下：

|  |  |  |
| --- | --- | --- |
| 模型 | 传统方法 | 神经网络 |
| hkust | 48.23% | 29.24% |

（中文模型）

|  |  |  |
| --- | --- | --- |
| 模型 | Chinglish测试集 | Librispeech测试集 |
| Chinglish | 21.01% | / |
| Chinglish + Librispeech  （传统方法） | 28.91% | 9.89% |
| Chinglish + Librispeech  （神经网络） | 15.31% | 4.44% |

（英语模型）

1. **关键词检索**

关键词检索的系统我在两个模型的基础上跑通过，一个是Babel系统使用的Cantonese数据自己训练的模型，另一个是CVTE的模型。其次，我还尝试在基于音素的网格上进行检索，也尝试过检索集外词[11]。总体效果总结如下：

|  |  |  |  |
| --- | --- | --- | --- |
| 系统 | Cantonese nnet2 | Cantonese Chain | CVTE模型 |
| CER | 58.63% | 48.63% | 7.07% |
| EER | 35.2% | 28.8% | 1.1% |

1. **关键词检索理论**

看了一些关键词检索领域的论文，并且在实验室的分享会上和大家分享了有关WFST和关键词检索领域的相关知识。

总结

最后的总结主要从两个方面来说，一个方面是技术层面，另一个是工作经历，个人成长层面。

首先，从技术层面来说，我入门了语音识别和关键词检索领域。这个领域当前变化非常快，需要不断更新，学习。但这之前，对语音识别，关键词，机器学习等的基础知识还是需要牢固地掌握。在实验室期间，在老师和师兄，同学的帮助下，我也开始接触了一些科研，论文的相关的东西。在这个方向上，我还需要多多学习。除此之外，计算机基本的工具的掌握也非常重要，这对于工作效率的提高也有很大的帮助。

这一段工作经历让我认识了很多人，也受到了老师，师兄和同学的很多关照和指引。有了他们的帮助，我能在未来的学习，工作中更好的前进。

参考文献

[1] D. Povey, Kaldi Document, pages on <http://kaldi-asr.org/doc/>

[2] Takaaki Hori Atsushi Nakamura, "Speech Recognition Algorithms Using Weighted Finite-State Transducers" [Synthesis Lectures on Speech and Audio Processing](http://ieeexplore.ieee.org/xpl/seriesAvailBooks.jsp?publisher=Morgan+&+Claypool&series=40)

[3] Mohri, Pereira and Riley, ["Speech Recognition with Weighted Finite-State Transducers"](http://www.cs.nyu.edu/~mohri/pub/hbka.pdf)Springer Handbook on SpeechProcessing and Speech Communication, 2008

[4] G.David forney, JR, "The Viterbi Algorithm" Proceedings of the IEEE

[5] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques" JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617

[6] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb 1989.

[7] Daniel Jurafsky & James H. Martin, "Chapter 4 N-gram" Speech and Language Processing

[8] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky and O. Yilmaz, "Quantifying the value of pronunciation lexicons for keyword search in lowresource languages," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, 2013, pp. 8560-8564.

[9] D. Povey et al., "Generating exact lattices in the WFST framework," 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012, pp. 4213-4216.

[10] D. Can and M. Saraclar, "Lattice Indexing for Spoken Term Detection," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 8, pp. 2338-2347, Nov. 2011.

[11] G. Chen, O. Yilmaz, J. Trmal, D. Povey and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, 2013, pp. 416-421.