基于深度学习的中文语音识别的建模研究

张　民
惠州学院网络与信息中心　广东惠州　516007

摘　要：随着科技的发展，语音识别的应用越来越广泛，智能化的语音识别有着非常重要的意义。文中对语音识别系统的工作机制、分类进行介绍，设计系统开发环境和框架。对基于深度学习的中文语音识别中的语音数据集收集、语音数据预处理、语音数据特征提取、构建声学模型、构建语言模型进行设计。该研究能够自录语音或者上传语音到服务器进行中文识别，支持将已识别出的中文翻译成英文的功能等。该研究能够为后续的语音识别的深入研究奠定基础。

关键词：深度学习；语音识别；特征提取；DFSMN模型

引言：

语音识别技术，也被称为自动语音识别（Automatic Speech Recognition，ASR），它已经渐渐与人们的生活密切相关。语音识别的目标是将人们说话的声音转化成计算机可以理解的二进制语言再进行相应的处理。与文本分类、机器翻译一样，语音识别是人工智能中自然语言处理（Natural Language Processing，以下简称NLP）的一个子领域，在人工智能非常热门的时代，从Siri到小度，从小冰到小娜，再到小爱同学，这些智能语音助手正在融入人们的生活[1]。语音识别技术的应用领域非常广泛，有智能家居、移动设备、智能客服、车载系统、智能医疗、工业控制、智能玩具等，它的核心就是通过语音与机器进行交互，让机器完成相关的任务。

1　语音识别系统

1.1语音识别系统工作机制

语音识别的任务是将语音序列转换为文本序列。有两种语音识别转换方式，一种是将语音直接转换为文本，一种是将语音先转换为音素（或者拼音），继而转换为文本。由于中文同音字较多，同样的音能表示不同的字，这就很考验输入语音的上下文语境，在训练出文本结果的同时还要考虑到上下文衔接，给训练大大增加了难度，将训练的精力分散开，最终造成识别准确率低的问题，因此第一种方法的可行性不高。对比第一种识别方法，第二种就能够合理分配训练内容，也就是只训练语音序列转化为音素，只训练音素转化为文本，这样语音识别率就能大大提高，文中的语音识别设计将会围绕第二种方式展开[2-3]。

1.2语音识别系统的分类

语音识别技术可根据不同的应用场景分为三大类：限制用户的说话方式，限制用户用词范围，限制系统用户对象[4-5]。

限制用户的说话方式。按照语音识别系统对用户说话方式的限制，可以分为孤立词识别系统，连续语音识别系统，即兴口语语音识别系统。连续语音识别系统是指中大规模词汇但是使用子词作为基本识别单元的识别系统；由于输入是随机的，因此语音内容也是随机，伴随着不少的随机事件，如吞咽、断续、结巴、重复、犹豫、咳嗽、喘气等，这些特点使得即兴口语语音识别充满挑战。限制用户用词范围。用词范围可以分为三种：小词汇量，中等词汇量，大词汇量和无限词汇量。

1.3语音识别的主要问题

（1）人为因素：不同的人说话方式、口音、语速、音量不同；同一个人的说话方式又会根据说话人的情感变化和身体状况而发生改变，生气的人说话语音节奏快，音调升高，音量大而生病的人说话节奏慢且音量较少，每个人的说话方式会随着时间变化，这些情况都会导致语音识别率的降低。

（2）环境噪音：实际的录音场景经常会有不同的环境杂音噪音，车鸣笛，风雨声，白噪音，旁人说话的声音等杂音噪音被录进音频当对语音识别有严重影响，导致识别率降低。

（3）硬件因素：不同的录音设备由于录音的性能不同，录音的方式不同采集的音频的采样频率语音信号都有所区别，进而影响语音的正确识别。

（4）语音识别系统对语义的理解：人类语音有词汇、同音字等不同形式，同样的发音有时候根据不同的上下文对应的词汇不同，因此要建立一个理解语义的规则。

2　基于深度学习的中文语音识别框架设计

2.1整体架构设计

基于深度学习的中文语音识别架构分为两部分，声

学模型训练和模型使用。声学模型训练首先需要收集数据集，接着对数据集做预处理，而后将语音数据一条一条做特征提取处理后得到特征向量，输入到声学模型，利用误差反向传播算法（Error Back Propagation，简称BP算法）求出神经网络的参数，最后训练出比较理想的声学损失较小的模型。训练完得到声学模型后，可以自行输入语音数据到训练好的声学模型里，声学模型得出拼音识别结果后，将拼音识别结果输入到语言模型中，语言模型根据简单的词频统计计算出最有可能的词语组合，查询词频字典后将拼音转为汉字，最后将整个文本进行输出。中文语音识别系统的架构如图1所示。

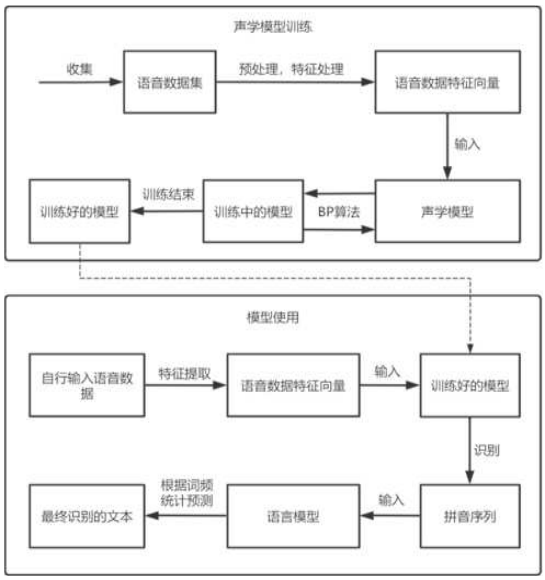

图1　中文语音识别系统的框架结构

3　语音数据的收集与处理

3.1 语音数据收集

文中所涉及的语音，包括语音数据集，设备录音音频，自行上传的音频等的采样率为16kHz。收集语音数据集是做语音识别的第一步，很多商业化的语音识别数据集并不对公众开发，尽管如此，也有一些公开可用的高质量数据集供训练。常见的中文语音数据集有THCHS-30、AISHELL、Magicdata、Primewords Chinese Corpus Set 1、Aidatatang_200zhST-CMDS等，这六个数据集共计约1385个小时。由于计算机资源有限，本次中文语音识别系统选择使用THCHS-30、ST-CMDS和Primewords这三个数据集来训练声学模型。

3.2 语音数据预处理

语音数据预处理主要是对数据集的划分和标定标签数据这两部分内容。在准备语音数据的同时，需要对语音数据进行标定标签数据。每一次的训练结果都要与正确结果做比对，利用反向传播算法把损失函数的误差从输出层向隐层一直到输入层逐层反向回传，把误差分摊给各层的单元，从而达到更新求解权重值W和偏置值b的目的。在训练声学模型输入语音数据时输入txt文本即可，这里包含了标签，一条语音的文件目录对应着其相应的拼音序列和中文文本序列。在数据标签里面，一条数据引用两个txt标签文本，因此一共有12个txt标签文本，数据标签如图2所示。

```
20170001P00142A0070 ST-CMDS-20170001_1-OS/20170001P00142A0070.wav
20170001P00142A0070 mei2 qian2 yi4 fen1 qian2 mei2 you3 lao3 yue1 han4 ka3 li3
```

图2　原标签样例

文中改进标定标签数据的方法，具体形式为wav_data_path \t pinyin_list \t hanzi_list，如图3所示。可以看到标签里的拼音字符是有带音标的，但是声学模型输出的时候，实际上是生成一系列的序号，自己制作一个JSON字典包，也就是model\model_language\pinyin_dict.josn文件，负责将这些序号对应为结尾带数字的字母字符串，比如na4，ni3等，而不是自带音标的拼音字符。

```
ST-CMDS-20170001_1-OS/20170001P00444A0119.wav　nà ní gàn má qù yī yuàn 那你干嘛去医院
```

图3　文中标签样例

语音上面的文本只是表示一条语音数据的标签，有9个标签文件，分别为THCHS-30的thchs_train.txt、thchs_dev.txt、thchs_test.txt，ST-CMDS的stcmd_train.txt、stcmd_dev.txt、stcmd_test.txt和Prime的prime_train.txt、prime_dev.txt、prime_test.txt，每个txt文件放着对应的数据标签条数。WAV语音数据集文件与标签文本放在一起，在声学模型训练时，直接引入数据集根目录datasets即可。

3.3 语音数据特征处理

在训练某一内容的时候需要提取想要训练内容的特征，语音识别领域也不例外，先把语音特征提取出来，神经网络训练时就会根据这些特点去做判断和分类。以特征信息量来说，Fbanks多于MFCC，MFCC多了离散余弦变换（Discrete Cosine Transform，以下简称DCT）等步骤，计算量更大，这其实是对语音信息的损变，损失了大量的声音细节。本中文语音识别使用Fbank特征提取方法。

4　基于深度学习的中文语音识别的建模

4.1 构建声学模型

在自动语音识别系统中使用声学模型（Acoustic model）来表示音频信号和构成语音的音素或其他语言单位之间的关系。现代语音识别系统使用声学模型和语言模型来表示语音的统计特性。声学模型模拟了语言中处理后得到的语音特征和语音的函数关系。继而利用语言

模型将得到与给定音频片段相对应的顶级单词序列。声学模型的建立，是整个语音识别系统最为关键的一环，一个语音识别系统的好坏很大程度取决于系统中声学模型的好坏。大多数现代语音识别系统都是在小块的音频上运行，即帧，每帧的持续时间大约为10ms。可以利用mel-frequency倒频谱等特征提取方式对每一帧的原始音频信号进行变换。这种转换的系数通常被称为梅尔频率倒谱系数（MFCC）s，并与其他特征一起作为声学模型的输入。声学模型的音频可以使用不同采样率，声学模型训练时采用的采样率最好能够和被识别语音音频具有相同的采样率和位记录，这样才能获得最好的语音识别效果。

4.2 构建语言模型

构建基于统计的语言模型，需要先收集数据量足够大的词频统计文本，包括单音节词、双音节词等。利用概率统计学的方法构建语言模型，输入拼音序列，就能得到出现概率最大的汉字序列，再将其作为最合理的句子进行输出。在统计语言模型里，每一个词的出现仅考虑与前面的词有关。通常考虑与前一个词或者前两个词，就已经能得到足够高的准确率了，这分别称为统计一元语言模型和统计二元语言模型，在极少数的情况下，才考虑三元，四元等语言模型。但是元的级数越高，计算的时间复杂度就越高，当要处理的拼音文本比较长时，普通的计算机会计算得非常吃力，带来无法避免的时间成本。在本语言模型里收集了一元词和二元词的词频统计字典，数据量分别是6880条和568647条。

基于马尔科夫链，实现拼音向文本的转换。马尔科夫链是基于动态规划算法实现的，类似寻找最短路径的算法。汉字与拼音的匹配可以看成同音字与拼音的一种通信，从左到右做匹配，如图4所示。

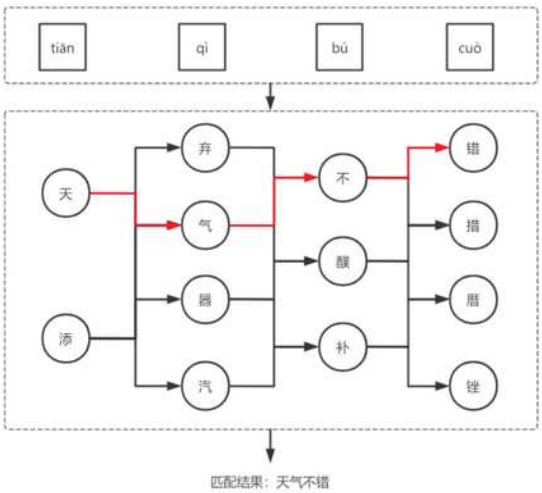

图4　拼音转汉字有向图

4.3 训练模型代码文件使用说明

训练模型的文件夹“modelCode”，各文件内容分别如下：

cnn_dfsmn_ctc：存放已训练好的声学模型。

datasets：存放语音数据集和标签文件。

model：包含model_language文件夹，声学模型AcousticModel.py和LanguageModel.py，model_language文件夹里声学模型和语言模型需要用到的文本，包括拼音序列pinyin_dict.json，单字词频统计文本language_word1.txt，双字词频统计文本language_word2.txt，拼音字典dict.txt。

plain.py：包含绘制时域图，频域图，频谱图的函数。

train_and_test.py：包含训练声学模型和加载声学模型的函数。可通过调用加载声学模型函数来做语音识别测试。

wav_speech_recorder.py：包含录制wav语音的函数。

5　结束语

文中使用能够对序列前后依赖关系建模的DFSMN框架，提高了模型的识别率。论文首先介绍了语音识别系统，说明了语音识别系统的工作机制、分类，然后设计了系统开发环境以及整体的框架。最后详细描述了设计过程中的步骤的方法，分别是语音数据集收集、语音数据预处理、语音数据特征提取、构建声学模型、构建语言模型。从应用角度看，该应用可以实现在WEB端自录语音或者上传语音到服务器进行中文识别的功能，支持将已识别出的中文翻译成英文的功能等。从研究角度上看，虽然语音识别技术涉及的学科和技术非常复杂，但是对于整体架构来说，包括数据集收集、特征提取、声学模型框架选择和神经网络的设计以及语言模型的建立，是比较合理科学的，能够为后续的深入研究奠定基础。

参考文献：

[1]王阳，牛长流，马国昊，王乐，牛青妍，王天.基于深度学习的多人语音识别研究[J].数字技术与应用，2021，39（07）：71-74.

[2]张丹.深度学习神经网络在语音识别中的应用探讨[J].电子世界，2021（06）：67-68.

[3]张允耀.基于深度神经网络的鲁棒性语音识别系统设计与实现[D].青海师范大学，2021.

[4]冯天艺.基于多任务神经网络的多维语音识别技术研究[D].南京邮电大学，2020.

[5]白璐，王连明.基于卷积神经网络的大容量汉语孤立字语音识别方法[J].东北师大学报（自然科学版），2020，52（02）：52-57.