文章编号:1007-1423(2019)31-0006-05 DOI:10.3969/j.issn.1007-1423.2019.31.002

基于 RNN 的桂柳方言语音识别系统研究

杨波

(武警广西总队参谋部,南宁 530031)

摘要:
随着人工智能技术的发展,语音识别也成为一个热门的研究方向。桂柳话作为广西壮族自治区的主体民族语言,其语音识别研究具有重要的现实意义。通过搭建以 RNN 为声学模型的语音识别系统,并应用联结时序分类准则于声学模型训练之中,经过大量的数据测试,该系统的语音识别正确率可达 92.7%,有着广阔的社会应用价值。

关键词:
桂柳话; 语音识别; 循环神经网络; 声学模型

0 引言

语音识别是指将语音自动转换为文字的过程。在实际应用中,语音识别通常与自然语言理解、自然语言生成及语音合成等技术相结合,提供一个基于语音的自然流畅的人机交互系统。语音识别技术的研究始于 20 世纪 50 年代初期,迄今为止已有六十多年的历史。1952 年,贝尔实验室研制了世界上第一个能识别十个英文数字的识别系统。20 世纪 80 年代,基于马尔科夫的建模方法推动了语音识别技术的蓬勃发展。近十年来,语音识别的发展又取得了长足的进步,国内外许多科研机构研发出了各自的语音识别系统,如微软、科大讯飞、捷通华声等。2011 年,微软的俞栋等人将深度神经网络成功应用于语音识别任务中,在公共数据上词错误率相对降低了 30%。

桂柳话是广西地区方言文化,系属西南官话的一种,是广西壮族自治区通行最广的汉语方言。作为面向东盟的前沿与窗口,广西已先后成功承办了 16 届中国东盟博览会,通过深化东盟国际合作,精耕细作加速融入“一带一路”建设。在广西加快建设中国-东盟信息港大数据中心的形势下,语音识别等人工智能技术的研究应用,必将为新型智慧城市创新、网络视听产业基地建设等打下良好基础。

1 RNN 的概念及应用

循环神经网络(Recurrent Neural Networks,RNN)因其循环递归处理历史数据和对历史记忆进行建模的特殊特性,适用于处理时间、空间序列上有强关联的信息。循环神经网络是深度学习中的一个重要分支,近年来循环神经网络模型相关的研究发展迅速。其中的成功案例包括手写字体识别、语音识别、自然语言处理和基于计算机视觉等序列问题。从生物神经学角度看循环神经网络,可以认为其是对生物神经系统环式链接的简单模拟,而这种环式链接在新大脑皮质中是普遍存在的。这也从侧面反映人类学习是一个动态变化的过程,因而对神经元的模拟在生物工程上有着重要的意义。

循环神经网络模型通过用于描述动态的序列数据,随着时间的变化而动态调整自身的网络状态,并不断循环传递,还可以接受广泛的序列信息结构作为输入。不同于前馈神经网络(例如 ANN、DNN、CNN 等),循环神经网络模型更加重视网络中的反馈作用。由于存在着当前状态与过去状态或者与未来状态的链接,循环神经网络模型可以具有一定的记忆功能。普通的深度神经网络是从左到右逐层传递的,其网络的神经元数据不断向前传递直到输出,所在层(当前层)的神经元之间并没有连接关系;而循环神经网络不同于前

馈式的神经网络，其引入了定向循环机制，神经元之间互相依赖、互相连接，因此能够处理前后关联的序列数据。

序列数据也可以被称为“序列信号”，而序列信号几乎无处不在，只要有先后关联关系或者时间关系的信号数据，都可以被认为是序列数据。在我们生活的时间和空间里，身边所发生的所有变化都可以使用序列数据来表示。如路由器根据访问网络的地址信息不断地调整自身所携带的信息；淘宝会根据用户点击商品的顺序，推测出其可能购买的商品，进而推荐相应的商品广告等，都是应用序列数据的例子。正是因为序列数据无处不在，与我们的日常生活息息相关，所以对序列数据建模显得十分重要。循环神经网络模型在语音识别中有着重要的应用，如使用双向循环神经网络模型输入音频数据，可以快速预测其对应的词组，其准确率可达到 90%以上。另外，使用双向循环神经网络模型实现单通道音乐的人声分离，实验结果表明该双向循环神经网络模型能够正确地从单通道的歌曲中分离出人声和背景音乐，该技术可以应用在手机麦克风中，在嘈杂环境下过滤掉背景噪声并提取出音频信号中的原声。

2 语音识别系统设计

2.1 桂柳话语音特点

桂柳话通行于广西壮族自治区五十六个县市的县城及圩镇地区，细分又有桂林话、柳州话、郴州话、荔浦话、平乐话等，其间有一些差别，桂林话受湖南话的影响比较多，而柳州话受广东话、壮话的影响比较多。桂柳话主意系统排列为：声母 19 个，韵母 37 个，声调有 4 个，外加一个入声调，共五个。此外，桂柳方言没有明显的轻声和变调。桂柳话一般没有汉语拼音的卷舌音 zh(之)，ch(吃)，sh(师)，r(日)，分别以 z(资)，c(此)，s(斯)，y(一)代替；er(而)等音节以 e(俄)代替，明显特征是有鼻化韵、夹杂入声(喉塞音)塞音尾(广泛流行于其它地区的没有入声韵尾)。桂柳话存在大量合音现象，即将两个甚至多个音节快速连读合成一个音节。此现象使用频率较高，广泛存在于代词、副词以及语气词等常用词汇中。一般来说，合音词声母取自合音上字，韵母取自合音词下字，声调来源于上字或者下字。就发音而言，桂柳话与普通话的音调关系：第一声(阴平)在方言里一般仍是第一声，第二声(阳平)一般是第三声，第三声(上声)一般是第四声，第四声(去声)一般是第二声，没有明显的轻声，说话时几乎字字重读，连语气词也有相当确定的声调。

2.2 语音识别框架

自动语音识别(Automatic Speech Recognition，ASR)是人工智能的重要入口，是一种让机器通过识别和理解，把人类的语音信号转变为相应文本的技术过程。早在 20 世纪 90 年代初期，就已经出现众多语音识别领域的研究人员试图利用人工神经网络 ANN 进行自动语音识别方面的研究，可是大部分效果并不理想，原因主要有：语音数据有限、神经网络容易过拟合、计算资源有限等。而与此同时，基于概率论的技术在语音识别领域得到蓬勃发展，例如高斯混合模型(Gaussian Mixture Model，GMM)、隐马尔科夫模型(Hidden Markov Model，HMM)等。语音识别技术已经出现了 20 多年，为何近年来才成为人工智能的主流技术呢？这要得益于深度学习技术，将语音识别领域的准确率提高到足以应用于实际环境中。自动语音识别技术提炼优化为一个框架结构，该模型主要分为编码(Encoder)和解码(Decoder)阶段，如图 1 所示。

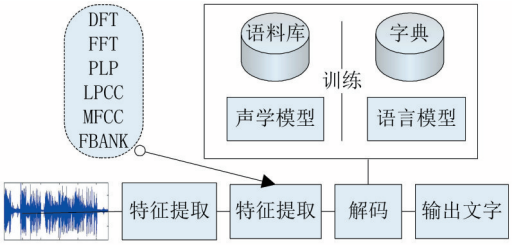

图 1 语音识别系统框架

其中，编码是将音频数据作为输入，转换成音频向量数据；特征提取是通过算法或者音频特征算法提取音频向量，提取后的特征为“声纹”，例如使用快速傅立叶变换(Fast Fourier Transform，FFT)对音频数据进行时域和频域间的转换；训练是从声纹数据和字典中学习固定特征，用于生成声学模型(Acoustic Model)和语言模型(Language Model)，声学模型用于识别语音向量，一般可以使用 GMM 或者循环神经网络等方法来识别向量，用 HMM 或者 CTC 来对齐输出的结果，语言模型是根据语法、语义规则对声学模型调整输出的结果，例

如修改与调整不符合逻辑规则的词语;在语音识别领域中,大部分模型并不是以单词作为基本单位,而是以音素作为基本的语音识别单位,音素是语音中最小的单位,依据音节里的发音动作来分析,一个动作构成一个音素,音素分为元音和辅音两大类,英语辅音和元音在语言中的作用就相当于汉语中的声母和韵母;解码是将训练好的声学模型和语言模型进行组合,输入新的声纹特征,最终输出其对应的文本文字。

使用深度学习技术进行语音识别,可以实现一个简单的声学模型,从简单的音频数据开始,对其特征提取得到“声纹”,通过循环神经网络模型实现一个声学模型,最后解码输出该音频数据所对应的文本文字。不同的人会有不同的语速,说话方式和行为也会不一样。例如,一个人可能会带有疑问地说出“HEEEEEELLO?”,而另外一个人则可能很开心地说出“HELLOOOOOOOOOO!”,这样对应同一个单词会产生不同长度的声音文件。而语音识别的任务就是把上面两个声音文件都正确地识别为“HELLO”。把各种不同长度的音频文件自动对齐到一个固定长度的文本是一件很困难的事情,循环神经网络帮我们很好地解决了这一难题,它能在特征提取阶段或是输出阶段对音素进行对齐操作。

2.3 基于RNN的声学模型

声学模型承载着声学特征与建模单元之间的映射关系。在训练声学模型之前需要选取建模单元,建模单元可以是音素、章节、词语等,其单元粒度依次增加。若采用词语作为建模单元,每个词语的长度不等,从而导致声学建模缺少灵活性;此外,由于词语的粒度较大,很难充分训练基于词语的模型,因此一般不采用词语作为建模单元。相比之下,词语中包含的音素是确定且有限的,利用大量的训练数据可以充分训练基于音素的模型,因此目前大多数声学模型一般采用音素作为建模单元。语音中存在协同发音的现象,即音素是上下文相关的,故一般采用三音素进行声学建模。由于三音素的数量庞大,若训练数据有限,那么部分音素可能会存在训练不充分的问题,为了解决此问题,我们采用决策树对三音素进行聚类以减少三音素的数目。

基于深度神经网络的声学模型是指用深度神经网络模型替换高斯混合模型,深度神经网络模型可以是深度循环神经网络和深度卷积网络等。该模型的建模单元为聚类后的三音素状态,模型如图 2 所示。图中,神经网络用来估计观察特征(语音特征)的观测概率和语音信号的动态变化(即状态间的转移概率)。Sn 代表音素状态;hM 代表第 M 个隐层。与基于高斯混合模型的声学模型相比,这种基于深度神经网络的声学模型具有两方面的优势:一是深度神经网络能利用语音特征的上下文信息;二是深度神经网络能学习非线性的更高层次特征表达。所以,基于深度神经网络的声学模型性能显著超越高斯混合模型的声学模型,成为当前主流的声学建模技术。

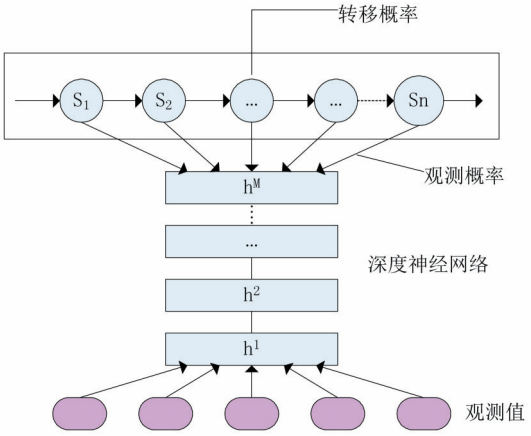

图2 基于RNN的声学模型

循环神经网络模型拥有记忆功能,用于影响未来时间序列的输出。首先把音频分成每份 20ms 长的音频块,即对应音频上的一帧数据。假设以每秒 16000 次的采样频率,那么一个 20ms 的音频对应 320 个采样数据。虽然只有短短的 20ms,但即使较短的音频片段也是由不同频率的声音交织而成,其中包括低音、中音和高音。为了使得音频数据更加容易地被循环神经网络处理,我们把一段连续的音频声波分解成很多段短暂的音频采样片段,例如刚才所说的 20ms 为间隔对音频进行切片采样。深度神经网络的输入是以 20ms 为单位的一帧,每一帧作为一个时间序列,使用音频文件经过声学模型的前馈计算,可以得到每一帧音频对应的汉字。循环神经网络模型使用 3 层的 GRU 网络模型,部分代码如下:

```
Def gru_model(input_dim=161,output_dim=29,recur_layers=3,
nodes=1024):
```

```
# RNN 层
For I in range(recur_layers):
# GRU 层
Output=GRU (nodes, activation='relu' , kernel_initializer=ini-
tialization,Return_sequences=True,name='rnn_{}'.format(i+1)
(output)
# 输出层(Softmax)
Time_dense=TimeDistributed(Dense(output_dim))(output)
```

2.4 系统识别训练

构建语音识别框架、准备语音数据、提取语音特征、建立声学模型之后，就要对声学模型进行训练了。语音识别声学模型的训练属于监督学习，需要知道每一帧对应的 label 标签才能进行有效的训练。在传统的语音识别声学模型中，在对语音模型进行训练之前，往往要求语音与文本进行严格的对齐操作，但它实际并不是一种严格的对齐方式，而是一种较为宽松的对齐方式。本文设计的语音识别系统，则是让深度神经网络自己去学习对齐的方式，从而引入了连接时序分类(Connectionist Temporal Classification，CTC)，CTC 层通过计算，使得输入与输出对应起来，减少了大量的标注时间，并使得声学模型能够做到端到端的有效训练。CTC 借用了 HMM 中的向前向后算法来计算可能路径，向前因子α和向后因子β定义为：

$$\alpha(t,u)=\sum_{\pi\in V(t,u)}\prod_{i=1}^{t}y_{\pi_i}^{i}$$

$$\beta(t,u)=\sum_{\pi\in W(t,u)}\prod_{i=1}^{T-t}y_{\pi_i}^{t+i}$$

向前向后算法通过动态规划的思想来解决，其针对一个当前标签 u 的全部路径累加，被分解为以 u 为前缀的全部路径的迭代累加，该迭代通过递归计算向前向后因子求得。经过高强度训练的声学模型，便可应用于系统测试。语音识别系统测试界面如图 3 所示。

图3 系统测试界面

3 结语

我国地大物博、人口众多，属于多民族国家，各地的方言也是多种多样，研究基于地方方言的语音识别系统具有重要的现实意义。本文在深入分析研究语音识别技术及桂柳方言音频特征的基础上，利用 RNN 的深入学习技术，构建完整的桂柳方言语音识别系统。测试结果表明，通过特征提取及模型训练，该系统在室内环境中对桂柳语音样本测试的识别率可达 92.7%，可以有效地在实际生活中桂柳方言对话场景进行应用，使人工智能技术能真正地服务于社会。

参考文献：

[1]Xue S，Yan Z. Improving Latency-Controlled BLSTM Acoustic Models for Online Speech Recognition[C]. IEEE International Conference on Acoustics，Speech and Signal Processing. IEEE，2017:5340-5344.

[2]蔡文俊. 广西柳州方言合音现象探究[J]. 文学教育，2013(18):126-129.

[3]张策，韦鹏程，石熙. 小语料库重庆话语音识别的研究[J]. 计算机测量与控制，2018(11):252-255.

[4]包晓安，徐海，张娜. 基于深度学习的语音识别模型及其在智能家居中的应用[J]. 浙江理工大学学报，2019(2):217-223.

[5]李德毅，于剑. 人工智能导论[M]. 北京:中国科学技术出版社，2018:111-214.

[6]陈仲铭，彭凌西. 深度学习原理与实践[M]. 北京:人民邮电出版社，2018:231-323.

作者简介：

杨波(1976-)，男，广西玉林人，硕士，工程师，研究方向为网络安全、人工智能、音视频信息处理

收稿日期：2019-09-24　　修稿日期：2019-10-18

(下转第 14 页)

参考文献：
[1]韩华瑞，代侦勇. 湖北省微博签到活动空间差异分析——以新浪微博为例. 测绘与空间地理信息，2016，39(10):159-162+166.
[2]杜翔，蔡燕，兰小机. 基于 Python 的新浪微博位置数据获取方法研究. 江西理工大学学报，2018，39(05):90-96.
[3]闵建. 基于签到数据的餐厅推荐技术研究. 杭州电子科技大学，2016:71.
[4]王丽鲲. 基于社交媒体地理数据挖掘的游客时空行为分析. 上海师范大学，2017:71.

作者简介：
白刚(1981-)，男，河北灵寿人，高级系统分析师，研究方向主要有地理信息、旅游信息化等
收稿日期：2019-09-20　　修稿日期：2019-10-31

Tourism Interest Point Mining Based on Tourist Check-in Data

BAI Gang

(School of Tourism Management, Guilin Tourism University, Guilin 541006)

Abstract:
Due to the convenience of GPS and LBS services, check-in point data with geographic data is ubiquitous. Integrates the characteristics of check-in data and geographic data, applies data cleaning, sorting and spatial analysis, and combined with the geographical characteristics of cities and traffic, so as to screen out the tourist spots with high interest from the massive check-in data with geographic information and provide decision assistance for tourists.
Keywords:
Data Cleaning; LBS; Spatial Analysis; Algorithm

(上接第 9 页)

Research on the Speech Recognition System of Guiliu Dialect Based on RNN

YANG Bo

(Chinese People's Armed Police Force Guangxi Unit Staff, Nanning 530031)

Abstract:
With progress of Artificial Intelligence technology, speech recognition has become a hot research orientation. Guiliu dialect is the main national language of Guangxi, recognition research of Guiliu dialect has great actual significance. Constructs a speech recognition using RNN as the acoustic model, and the training criterion based on Connectionist Temporal Classification is successfully applied to the acoustic model training. Through a large number of data testing, the recognition accuracy rate is up to 92.7%, so it has broader social application value.
Keywords:
Guiliu Dialect; Speech Recognition; Recurrent Neural Networks; Acoustic Model