基于 Kaldi 的语音识别

王 凯[1],马明栋[2]
(1. 南京邮电大学 通信与信息工程学院,江苏 南京 210003;
2. 南京邮电大学 地理与生物信息学院,江苏 南京 210003)

摘 要:人工智能技术是当前计算机科学的研究热点,人机通信是人工智能技术的重要组成之一。作为人机通信主要方法之一的语音交互也一直是科学家的研究热点,语音交互技术的关键是语音识别。而目前大多语音识别软件要么功能单一,要么价格昂贵,Kaldi 作为新兴的开源语音识别工具,凭借其强大的功能和简单的获取渠道逐渐流行。该文介绍了语音识别技术的发展历程,Kadli 软件的基本架构和其所具有的独特优势,语音识别的一般处理流程,多层神经网络的基本结构以及多层神经网络在语音识别当中的应用。对基于 Kaldi 软件当中的 HMM-DNN 模型,使用中文数据集训练该模型,搭建一个完整的语音识别系统。通过该系统,不仅能展现出 Kaldi 软件丰富强大的功能,同时也为语音识别研究人员选择合适的工具提供了新的思路。

关键词:人机通信;语音识别;Kaldi;多层神经网络;HMM-DNN
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2021)01-0013-05
doi:10.3969/j.issn.1673-629X.2021.01.003

Speech Recognition Based on Kaldi

WANG Kai[1], MA Ming-dong[2]
(1. School of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. School of Geographical and Biological Information, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Artificial intelligence technology is the current research hotspot of computer science, and human-machine communication is one of the important components of artificial intelligence technology. As one of the main methods of human-computer communication, speech interaction has always been a hot topic among scientists. The key of speech interaction technology is speech recognition. The current speech recognition software is either single-function or expensive. As an emerging open source speech recognition tool, Kaldi is gradually popular with its powerful functions and simple access channels. We describe the development of speech recognition technology, the basic architecture of Kaldi software and its unique advantages, the general processing flow of speech recognition, the basic structure of multi-layer neural networks and the application of multi-layer neural networks in speech recognition. The HMM-DNN model in Kaldi software is trained by Chinese data sets, and a complete speech recognition system can be built. This system not only shows the rich and powerful functions of Kaldi software, but also provides a new idea for speech recognition researchers to select the right tool.

Key words: man-machine communication; speech recognition; Kaldi; multi-layer neural network; HMM-DNN

0 引 言

随着科技和经济的快速发展,互联网和人工智能技术对生活影响越来越大。人们可以通过键盘、手写输入等方式与计算机进行通信,但是为了更加智能地与计算机进行通信,语音通信显然是更好的方法,而语音通信的关键技术就是语音识别[1]。

语音识别最早起源于1952年贝尔研究所的仅能识别10个英文数字发声的实验系统,而后的近二十年时间,语音识别向大规模方向发展,但是识别的内容还是很局限[2]。到了八十年代之后,科学家对语音识别技术的研究方法发生了很大的变化,由于基于统计模型的语音识别理论的成熟,传统的标准模板匹配技术

收稿日期:2020-03-09 **修回日期**:2020-07-13
基金项目:江苏省自然科学基金-青年基金项目(BK20140868)
作者简介:王 凯(1996-),男,硕士研究生,CCF 会员(B7678G),研究方向为数字图像处理;马明栋,博士,教授,研究方向为地理信息系统平台软件设计与开发等。

逐渐被这种新兴的方法所取代。同时为了适应时代的发展,小词汇量显然不能够满足需求,因此大词汇量的研究由此开始,对于识别语音的来源也不仅仅局限于某个特定人的语音,非特定人的连续语音识别成为研究重点。

语音识别开始于美国的六年后进入中国,1956 是国内研究语音识别技术的起点,但当时社会的科技水平和经济条件不足以支持该项技术的研究。因此直到国家开展了 863 计划后,有了充分的经济和硬件条件的支持,语音识别技术才重新进入了科学家的视野,也从此开始逐渐走上正轨。虽然研究时间不是很长,但是取得的研究成果已经惠及了生活的方方面面,像语音输入法、智能机器人、智能家居等等[3]。

1 Kaldi 简介

Kaldi 是一个非常强大的语音识别工具,其目前的支持和维护工作主要由 Daniel Povey 负责。Kaldi 的代码是开源的,并且目前仍然在 Github 上非常活跃。该软件不仅整合了另一款语音识别软件 HTK 的基本功能,还在其基础上加入了深度神经网络分类器(DNN)[4]。

对于语音研究者们而言有许多语音识别工具可以使用,像上述提到的使用 C 语言编写的 HTK,使用 java 语言编写的 Sphinx-4 等[5]。但是 Kaldi 具有的以下独特优势使其不断被开发和应用:

· 集成了有限状态转换器(FST),让其作为一个库方便开发者进行使用。

· 支持更加广泛的线性代数运算,在其矩阵函数库中包含 BLAS 和 LAPACK 两种运算。

· 软件和算法的设计采用通用结构,方便开发者根据需求进行修改和拓展。

· 软件的代码是开源的,开发者可以修改代码并重新发布。

Kaldi 的核心代码是由 C++编写,在此之上使用 bash、python 等脚本语言写了一些上层的工具,软件的结构如图 1 所示。图 1 中 Kaidi 的体系结构主要分为四个部分,首先是最上面的外部工具,包含上文提到的矩阵运算库等;接着是 Kaldi 库,该库里面主要包含的代码是 GMM 和 HMM 等多种训练模型的代码;然后是编译完成的可执行程序;最后一部分是脚本程序,用来实现对对语音训练步骤的控制[6]。

目前 Kaldi 软件的官方文档不像 HTK 官方文档那样有许多关于统计语音识别的基础介绍性文档,因此对于非专业语音识别研究人员或者是语音相关基础知识较薄弱的人来说,在使用该软件之前需要全面了解一下语音信号处理相关的基础知识。Kaldi 软件除了上文提到的几个独特优势之外,还有以下几个对开发者有利的特性。第一点是 Kaldi 的代码是经过全面测试的,从而保证了代码的正确性,不仅保证了使用软件的稳定性,同时方便开发者在发生错误时进行快速准确的定位。第二点是 Kaldi 的代码都比较易于理解,在了解了基本语音信号处理的相关知识后就可以快速地理解代码的含义,清楚代码中每个模块的作用。最后一点是 Kaldi 当中每个模块都比较小,各个模块之间的耦合性也比较小,因此其很方便修改和复用。

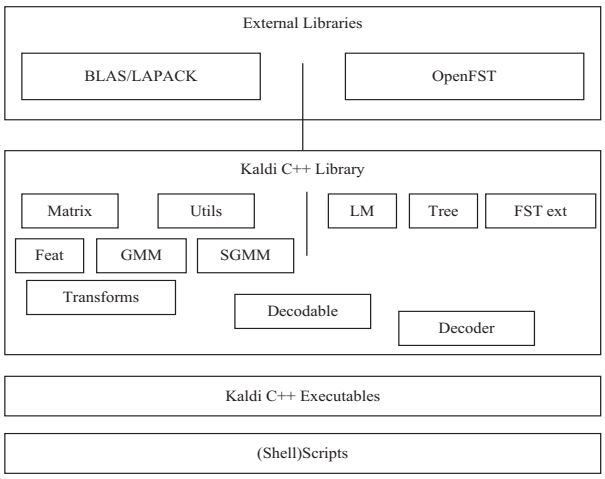

图 1 Kaldi 软件结构

2 语音识别过程

语音识别相对于机器翻译要更加得复杂和困难。机器翻译系统的输入一般是印刷文本内容,计算机可以较为清楚地分辨出输入当中的单词和单词串,然而语音识别系统的输入是人们的语音,输入内容的复杂度就增大了许多[7]。因为输入的语音内容具有很大的不确定性,现实生活中人们往往可以根据说话人的音调、手势、面部表情等来获取信息。而单一的语音输入使得计算机不具备这些条件去获得这些信息,因此想要使得计算机像人一样去准确地识别语音中的内容是十分困难的[8]。

语音识别过程主要包括以下几个步骤:

语音信号采集:信号的采集是进行语音信号处理的前提。声音通过麦克风输入计算机,此时麦克风所采集的声音信号是模拟信号,需要转换成计算机能够处理的数字信号。

语音信号预处理:对于采集到的信号,首先进行滤波操作,滤除语音信号当中所包含的噪声。接着进行 A/D 转换,将模拟语音信号转换成为计算机能够处理的数字信号[8]。然后进行预加重处理,提升语音信号的高频部分,平坦信号的频谱,便于进行频谱分析。预处理的最后一个步骤就是进行端点检测,检测该段语音信号中声音的起点和终点,提高信号的处理效率。

语音信号特征参数的提取:人说话的频率一般在 10 kHz 以下。因此根据香农采样定理,使用 20 kHz 的采样频率进行采样就不会造成频谱混叠。采样后可以使用多种方式来进行语音信号特征参数的提取,像线性预测系数(LPC)、感知线性预测(PLP)、梅尔频率倒谱系数(MFCC)等[9]。语音信号当中包含许多特征参数,不同特征参数构成的特征向量具有不同的声学和物理意义,上面提到的两种线性预测的方法主要是根据声管模型来提取参数,该方法提取的特征参数反映的是声道的响应特性。而目前使用较多的梅尔频率倒谱系数的特征参数提取方法则是根据人的听觉特性来提取特征参数,人类对于声音频率的感知是呈对数关系的,而经过梅尔倒谱系数的方法提取了特征参数后,人对于梅尔频率的感知是呈线性关系的。

语音识别:语音识别的方法有多种,可以分为两大类,第一类是传统模型。传统模型中包含基于动态时间规整算法(DTW),该种算法是目前连续语音识别的主流方法,算法的主要思想对于同一个单词不同的人发音的长短不同,从而声音信号产生的时间序列有较大的差异,通过将原始语音的时间序列进行延展,然后比较序列之间的相似性实现语音信号的识别。算法在技术上容易实现,并且识别的准确率也较高,缺点就是算法的运算量较大。基于参数模型的隐马尔可夫模型(HMM)的主要思想是根据观测的序列估计出想要得到的目标序列,因此为了能够提高识别的准确率需要使用大量的数据,数据越多用来推断的条件越充足。该方法主要用于大词汇量的语音识别系统,使用该算法时对计算机的内存要求较高,并且需要大量的训练数据和长时间的训练,因此算法也具有很高的识别准确率[10]。还有基于参数模型的矢量量化(VQ)方法,使用该方法所需的资源较上述的 HMM 方法都显著减少,但是该算法在大词汇量识别的准确率相较于 HMM 算法也有显著的下降,但其在孤立词识别的语音系统当中有很好的应用。

第二类就是现代的神经网络模型,也是该文将要使用的模型。该模型类似人脑,可以通过训练学习来达到较高的识别准确率[11]。一般情况下,语音识别都是通过分析语音信号得到信号的语谱图,再从语谱图获取信息来进行进一步的处理。语谱图一般具有结构性特点,这些特点会受到说话人以及说话人所处环境等因素的影响,因此考虑如何消除这些外在的因素从而更好地体现语谱图原有的特点。

卷积神经网络中提供时间和空间的平移不变性卷积,将该思想运用到语音信号的建模当中,利用卷积不变性的特点,避免了其他因素对信号特征的干扰。将得到的语谱图用图像处理过程中所使用的卷积神经网络进行处理识别,使用神经网络结构也方便对得到的信息进行处理运算,因为目前关于神经网络的相关框架也都比较成熟。

3　DNN-HMM 模型

首先对人工神经网络(ANN)而言,人工神经元是该网络的基本组成单元。网络有多个输入,经过每个神经元上的系数加权求和后输出,一个神经元模型可进行如下的描述[12]。首先是输入向量 $X_j = \{x_1, x_2, \cdots, x_n\}^T$, $x_i(1 \leqslant i \leqslant n)$ 为第 i 个神经元的输入, n 表示输入神经元的个数。$W_{ij}(1 \leqslant i \leqslant j)$ 是节点 i 和 j 之间的连接强度。b_j 为神经元 j 的阈值。使用 $x_0 = 1$ 的固定偏置输入节点表示阈值节点,与神经元之间的链接强度为 $-b_j$。

由上面的参数可得到神经元 j 的输出加权和为:

$$z_j = \sum_{i=0}^{n} x_i w_{ij} = \sum_{i=1}^{n} x_i w_{ij} - b_j \tag{1}$$

同时也可以求得神经元 j 的输出状态:

$$y_j = f(z_j) = f\left(\sum_{i=1}^{n} w_{ij} x_i - b_j\right) \tag{2}$$

其中,$f(\,)$ 函数为该神经元的激活函数,由该函数反映出神经元输入和输出之间的关系。该函数一般使用 Sigmod 函数,并且将该函数的输入限制在[0,1]之间,此时函数的表达式为:

$$y = f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}, x \in R \tag{3}$$

按照一定层次将单个的神经元连接起来,能够得到一个深层神经网络(DNN)[13-14]。一个多层神经网络模型如图 2 所示。

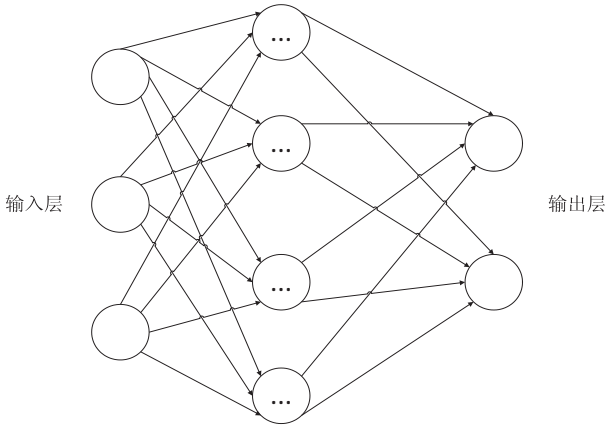

图 2　多层神经网络模型

一般的语音识别框架都是基于 GMM-HMM 的,然而该模型的层次较浅,对于数据间的深层特性,仅仅使用该模型的话无法进行捕捉[15]。而 DNN-HMM 模型则可利用 DNN 很强的学习能力来弥补该缺点,从而取得优于 GMM 模型的效果。下面给出了一个 DNN-HMM 系统的结构图(见图 3)。该结构当中

HMM 描述语音信号的动态变化,DNN 当中的每个节点的输出结果则被用来估计 HMM 某个状态的后验概率。

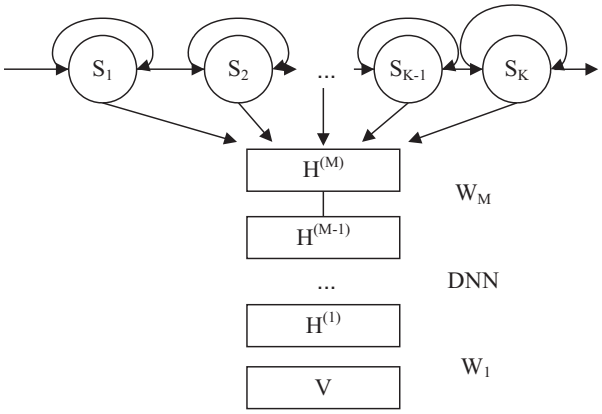

图 3 DNN-HMM 结构

若语音信号的声学输入为 $O=(o_1,o_2,\cdots,o_t)$,其为语音信号特征处理过程当中所获的声学特征向量[16]。句子为 $W=(w_1,w_2,\cdots,w_n)$,其由一系列的单词所组成。则进行语音识别的任务就是在给定声学输入的情况下得到最有可能的输出,可以用如下的式子表示:

$$W=\arg\max_W P(W\mid O) \tag{4}$$

如果要获得最好的识别结果,需要使 $P(W\mid O)$ 最大,则将上面公式继续展开可以得到:

$$\begin{aligned} W&=\arg\max_W P(W\mid O)=\\ &\arg\max_W p(O\mid W)P(W)/P(O)=\\ &\arg\max_W p(O\mid W)P(W) \end{aligned} \tag{5}$$

在一个语音识别系统中,$P(W)$ 为语言模型,$P(O\mid W)$ 为声学模型。设 $Q=\{q_1,q_2,\cdots,q_t\}$ 为一个状态转移序列,根据维特比解码算法,声学模型可进一步写为:

$$\begin{aligned} P(O\mid W)&=\sum_Q P(O,Q\mid W)P(Q\mid W)\approx\\ &\max\pi(q_0)\prod_{t=1}^{T}a_{q_{t-1}q_t}\prod_{t=1}^{T}p(o_t\mid q_t) \end{aligned} \tag{6}$$

其中,$a_{q_{t-1}q_t}$ 是状态 q_{t-1} 和 q_t 之间的转移概率。

在使用 DNN 时,DNN 能给出的只有输出层每个节点上状态的后验概率 $p(q_t\mid o_t)$,通过贝叶斯定理进行转换可得:

$$p(o_t\mid q_t)=p(q_t\mid o_t)p(o_t)/p(q_t) \tag{7}$$

4 实验过程和结果

Kaldi 软件可以运行在任何 Unix-like 的环境下,软件运行时对内存的要求较高,因此在使用该软件时要避免运行其他软件。本次实验需要训练的为 DNN 模型,为了数据集的快速训练,实验环境为一台带有 GPU 的 Ubantu 16.04 的机器,内存为 8 G。

有了实验环境之后,先安装 Kaldi 软件。如果计算机上安装了 git,可以直接使用以下命令从 Github 上下载该软件。

```
gitclone https;://github.com/kaldi-asr/kaldi.git kaldi -origin upstream
```

软件下载完成之后,需主要关注以下三个目录。./tools文件目录下存放的为 Kaldi 的依赖库,./src 目录下存放的为 Kaldi 的源代码,./egs 目录下存放的为 Kaldi 提供的一些例程。

然后进入 tools 目录下使用如下命令安装 Kaldi 所需要的依赖,依赖安装完成后编译源码。

```
cd kaldi/tools/
extras/check_dependencies.sh
make
```

等待源码编译完成之后,如果没有报错信息,就可以开始训练模型了。训练的数据集为 thchs30 中文公共数据集,该数据集可按照 README 文件当中的提示进行下载,下载后的文件解压到 egs/thchs30/s5 下面的新建文件夹 thchs30-openslr 下,该数据集主要包含 trian、dev 和 test 三个部分。

本次实验在单台机器上进行,因此对环境变量进行一些修改,并修改 s5 文件夹下面的 cmd.sh 脚本。数据集的存放位置也需要根据下载后的位置进行相应的修改。

```
export train_cmd="run.pl -mem8G"
export decode_cmd="run.pl -mem8G"
export mkgraph_cmd="run -mem12G"
```

以上的准备工作完成之后,就可以运行 s5 文件夹下面的./run.sh 文件。但如果直接执行 run.sh 脚本的话,可能会发生未知的错误,并且本次实验需要训练的只是 DNN 模型,其余的模型不需要训练,因此为了节约训练时间,应该单步执行./run.sh 下面命令。

模型训练完成后,即可进行识别的测试。但是直接使用麦克风或者是声音文件进行测试,识别的准确率很低。因此还需要安装扩展包来识别自身的语音,安装扩展的命令如下:

```
cd ./tools
./install_portaudio.sh
cd ../src
make ext
```

接着将声音文件的 online_demo 拷贝到 thchs30 文件夹下面,然后在该文件夹下新建 online-data 和 work 两个文件夹。online-data 文件夹底下再新建 audio 和 models 两个文件夹。audio 下面存放需要识别的语音文件,models 文件夹下面存放训练出来的模型 final.mdl 和有限状态机 HCLG.fst 等文件。

以上工作完成后将需要识别的语音放到/online-data/audio 文件夹下,返回上层文件夹,执行./run. sh 脚本,则识别开始识别音频文件里面的内容,并输出到命令行,输出如图 4 所示,识别的准确率约为 70% 。

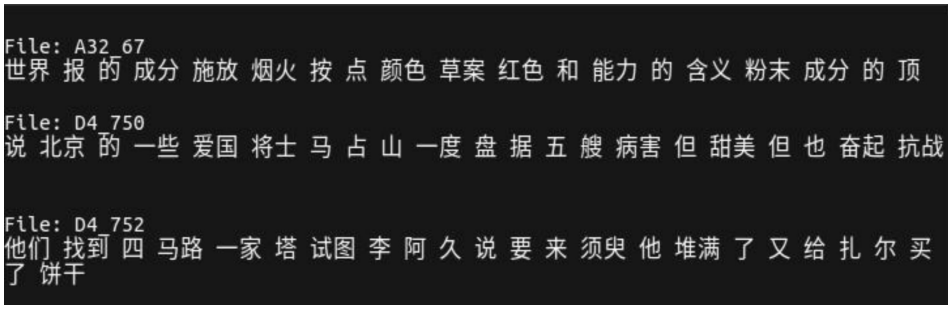

图 4　部分识别结果

5　结束语

介绍了语音识别的发展过程、语音信号处理的基本流程,重点关注语音识别软件 Kaldi,详细介绍了软件的体系结构,并且使用公共中文数据集训练了软件当中的 DNN 模型,实现了一个中文语音识别系统。搭建的识别系统虽然能够正常工作,但是识别的正确率并不高。导致该问题的原因之一是数据集的限制,还有一个可以改进的地方就是源代码中有关 DNN 结构的代码,改变或者优化网络的结构会提升识别的正确率。

参考文献:

[1] 朱春山. 基于 Kaldi 的语音识别的研究[D]. 南京:南京邮电大学,2018.

[2] 何湘智. 语音识别的研究与发展[J]. 计算机与现代化,2002(3):3-6.

[3] 胡文君,傅美君,潘文林. 基于 Kaldi 的普米语语音识别[J]. 计算机工程,2018,44(1):199-205.

[4] 杨胜捷,朱灏耘,冯天祥,等. 基于 Kaldi 的语音识别算法[J]. 电脑知识与技术,2019,15(2):163-166.

[5] 刘　琼. 几种开源英语识别工具包的对比分析[J]. 计算技术与自动化,2018,37(4):123-127.

[6] KIM Sang-Kyun, PARK Young-Jin, LEE Sangmin. Voice activity detection based on deep belief networks using likelihood ratio[J]. Journal of Central South University,2016,23(1):145-149.

[7] 赵　力. 语音信号处理[M]. 北京:机械工业出版社,2003.

[8] 吕　赫. 基于 DNN 的语言识别系统的研究与实现[D]. 成都:电子科技大学,2017.

[9] SUN R H, CHOL R J. Subspace Gaussian mixture based language modeling for large vocabulary continuous speech recognition[J]. Speech Communication,2020,117:21-27.

[10] 李敬阳,吴明辉,王　莉,等. 一种基于 GMM-DNN 的说话人确认方法[J]. 计算机应用与软件,2016,33(12):131-135.

[11] HINTON G E, OSINDERO S, YEE-WHYE T. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006,18(7):1527-1554.

[12] AL-RAHHAL M M, BAZI Y, AL-HICHRI H, et al. Deep learning approach for active classification of electrocardiogram signals[J]. Information Sciences,2016,345:340-354.

[13] 朱大奇,史　慧. 人工神经网络原理及应用[M]. 北京:科学出版社,2006.

[14] 朱　洁. 基于人工神经网络的信息安全加密管理评估[J]. 计算机技术与发展,2019, 29(9):97-101.

[15] 李鹏飞. 基于深度学习的维语语音识别研究[D]. 合肥:安徽大学,2016.

[16] ACHANTA S, GANGASHETTY S V. Deep Elman recurrent neural networks for statistical parametric speech synthesis[J]. Speech Communication,2017,93:31-42.