孤立词识别项目报告

姓名：张瑞阳
学号：17307130064
日期：2019年6月23日
指导老师：薛向阳教授，复旦大学计算机科学技术学院

1.背景

本项目实现了一个孤立词识别系统，本系统能够接收属于20个孤立词的语音，并预测出该语音所对应的孤立词。1952年，AT&TBell实验室的Davis等人研制了第一个能够识别十个英文数字的特定人语音识别系统Audry系统。本项目的目标就是达到上世纪50年代的语音识别技术水平。

1956年。美国普林斯顿大学RCA实验室的Olson和Belar等人研制出能够识别10个单音节词的语音识别系统，该系统采用带通滤波器组获得的频谱参数作为语音增强特征。19世纪60年代，苏联的Matin等人提出了语音结束点的端点检测技术，使语音识别水平明显上升。19世纪70年代初，科研人员提出了线性预测技术(LPC)和动态时间弯曲技术(DTW)，有效地解决了语音信号的特征提取和不定长语音的匹配问题，同时提出了矢量量化(VQ)和隐马尔可夫模型(HMM)。自此，语音识别领域进入快速发展阶段。

当前语音识别领域的最新趋势都和深度学习密不可分。当前最新的技术趋势是端到端技术(E2E)。端到端技术(E2E)建立了一种声学—单词模型，这种模型直接将单词作为输出单元。

连接时间分类(CTC)是端到端技术(E2E)的一种，但这种技术一直受到词典外词

(OOV)问题的困扰。当前该技术的最新进展是尝试改进声学—单词模型，这种改进形成了一种混合CTC模型，当基于单词的CTC模型在测试时间内发出OOV信号时，它会参考基于字母的CTC。然后通过训练混合单元CTC模型将所有OOV单词分解为频繁单词和多字母单元的序列。

其他端到端技术(E2E)和深度循环神经网络和深度卷积神经网络的强大的学习能力密不可分。有一种模型基于深度双向LSTM递归神经网络和连接时间分类目标函数，它能够将字符序列转录频谱图而无需中间语音表示。这个模型引入了对目标函数的修改，使得可以以最小化对任意转录损失函数的期望训练网络，从而即使没有词典或语言模型，这也可以直接降低单词错误率。

2.识别算法

2.1算法综述

本项目采用图像分类的方法解决孤立词识别问题。本项目首先对所有语音数据进行端点检测，去除无声段和噪声段，得到有声段语音数据。然后对语音数据利用短时傅里叶变换计算语谱图矩阵。因为此时的语谱图矩阵虽然频域维数相同，但是时域维数不同，不能输入到神经网络中进行计算，所以需要将语谱图矩阵变换成同一大小。本项目采用矩阵重新采样将语谱图矩阵变换成同一大小。得到数据集后，输入到卷积神经网络中进行训练，最终得到模型。系统框图见图1。

2.2端点检测

传统端点检测算法以整个语音的平均能量和平均过零率作为衡量标准，以此得到整个语音的有声段开始点和有声段结束点。这种做法虽然简洁有效，但是很难处理一段语音的有多段有声段的情况，比如在本项目解决的问题中，一个孤立词的两个字可能被分开说，这时候，这种基于短时平均能量和短时平均过零率的传统端点检测算法往往会提取出其中的一段有声段而裁剪掉了其他有声段。本项目的端点检测算法以帧为单位，采用了定长缓冲区的技术，有效地解决了上述问题。本项目的端点检测首先将语音分帧，然后定义了一个长度为固定帧数的缓冲区，并且实现了一个能够判断一个帧是否为有声帧的检测器。端点检测的工作流程如下，首先定义了一个标识符，用来表示当前帧是否已经进入了有声段。然后依次遍历所有的帧，当标识符为假时，当前帧会被直接存入定长缓冲区中，然后本端点检测算法会利用之前实现好的检测器统计当前定长缓冲区中有声帧的数量，并把它和 0.9 * 定长缓冲区的长度 进行比较，如果前者大于后者，本端点检测算法就判定当前语音帧已经进入了有声段，把标识符设为真，把当前定长缓冲区中所有帧都保存到有声帧列表中，然后清空定长缓冲区。当标识符为真时，当前帧会被直接存入有声帧列表中，同时存入定长缓冲区中，然后本端点检测算法同样会利用之前实现好的检测其统计当前定长缓冲区中有声帧的数量，并把它和 0.9 * 定长缓冲区的长度 进行比较，如果前者小于后者，本端点检测算法就判定当前语音帧又进入了无声段，同时会认为一

个有声段已经结束了，这时，它会把标识符设为假，返回有声帧列表，将定长缓冲区清空，同时将有声帧列表清空，然后继续遍历之后的帧。通过本端点检测算法，可以很好的解决一段语音中有多个有声段的情况，本端点检测算法理论上可以得到有任意多有声段语音的所有有声段，从而很好地解决了传统端点检测算法只能提取出一段有声段的缺陷。

2.3语谱图

即利用短时傅里叶变换计算计算语音的语谱图矩阵。语谱图的计算采用移动加窗，同时对每个窗取出的语音进行傅里叶变换即可。语谱图是语音数据的一个特征，同时语谱图和图像领域密不可分。

2.4图片大小变换

利用上一步计算出的语谱图矩阵生成的语谱图在频率域上维数相同，但是在时间域上长度不同，这是因为经过端点检测和有声段截取之后，每个语音的长度变成不是一样长的。但是训练神经网络需要输入的数据具有相同的大小，这就需要进行语谱图大小的变换，将语谱图时间域的长度变成一致的。通过导入图片处理库，可以将图片的大小通过重新采样的方式变换成相同的大小。在这里，重新采样能够在误差最小的条件下，将一个图片变换成另一个大小。在语音处理领域，有一些技术专门处理语音不定长问题，这些技术都是以误差最小为目标，将语音变换成同样的长度。在本项目中，将语音通过计算语谱图矩阵转化为语谱图，然后在图片的层面解决图片大小不一致问题，提供了解决语音处理领域语言不定长问题的一个新思路。

2.5卷积神经网络

卷积神经网络是解决图像分类问题的有力工具。卷积神经网络适合解决固定大小输入问题，卷积神经网络的本质是学习一个固定大小输入和固定大小输出之间的规律。在本项目中，通过制作数据集，提供大量输入和与之对应的输出，即可让卷积神经网络学习到输入和输出的规律。

卷积神经网络的工作流程其实十分简单，我们首先将卷积神经网络的参数初始化为初始值，然后对于训练集的一个输入和对应标签，我们先进行前向传播过程，让输入通过卷积神经网络得到卷积神经网络的预测值，这个预测值当然不可能是正确的，我们之后会按照一定的标准计算这个预测值和输入所对应的标签值之间的误差，之后我们进行反向传播过程，我们计算误差对所有参数的导数，实际上，误差对每个参数的导数值标示了每个参数该如何更新，从而使下一次预测的误差变得更小一些。之后我们就会对所有的参数进行更新。然后我们就完成了一个数据的训练过程。实际上，这个过程的本质就是实现了学习一个数据中的规律的目的。

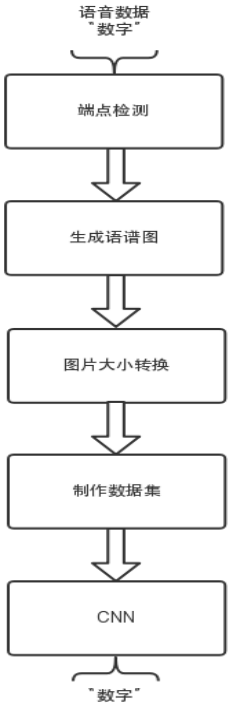

图1：系统框图，输入语音在通过端点检测之后，会被生成语谱图，语谱图经过图片大小转换变成固定大小，所有图片制作成数据集之后进行模型训练。

3.实验设置及数据集

3.1实验设置

3.1.1模型结构

模型采用经典的CNN4模型，CNN4模型有4个块，每个块有4个层，依次是conv2d, batchnorm, relu,maxpool2d。模型框图见图2、图3。

3.1.2训练参数

训练参数见表1。

参数	值
N_epoches	20
Batch_size	10
Loss_fn	CrossEntroyLoss
Optimiser	SGD
Learning_rate	0.001

表1：具体训练参数。

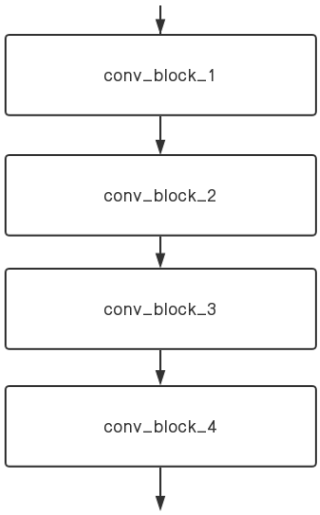

图2：*CNN4*模型结构，整个模型由*4*个*conv_block*组成。

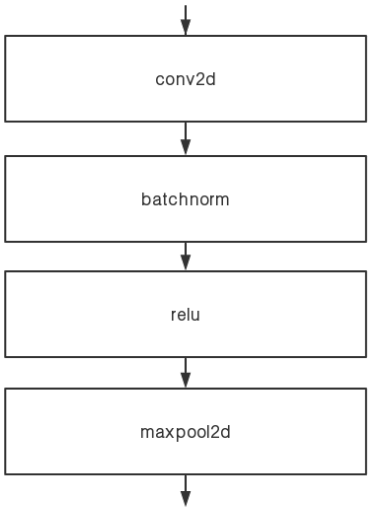

图3：*conv_block*结构，整个块由*4*个层组成。

3.2数据集

3.2.1所使用数据集介绍

本项目使用的数据集采用了30个人的语音数据。训练集和验针集的划分比例为4：1，即训练集包含24个人的语音数据，验证集包含6个人的语音数据。在使用语音数据之前，已经对所有语音数据进行了端点检测。

3.2.2原始数据分析

因为本项目是采用对语谱图进行图像分类来进行孤立词识别，所以十分关注语谱图的质量。在原始数据集中，通过初步分析发现，有语谱图会出现大面积空白的情况，如图4。通过大致的筛查和观察，排除了4个人的语音数据，最后剩

下30个人的语音数据。

在绘制语谱图的过程中，还发现了缺少数据的情况。在训练集中，本来应该出现9600个语谱图图片，最后发现只出现了9583个语谱图图片。在验证集中，本来应该出现2400个语谱图图片，最后发现只出现了2398个语谱图图片。通过输出错误信息最后发现，在端点检测之后，有的语音数据的长度变成0，即原始语音数据中没有语音。具体语音数据的信息如图5。

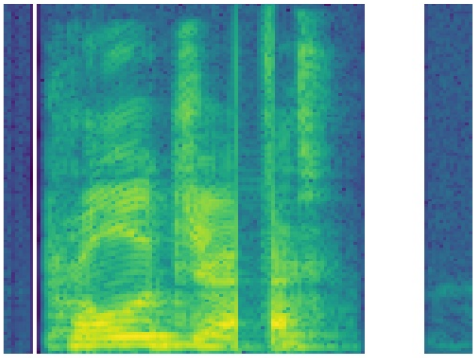

图4：语谱图出现大面积空白的情况的一个例子。

```
./tmpwav/trainset\15307130191-16-03.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\15307130191-16-05.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\15307130191-16-08.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\15307130191-16-11.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\15307130191-16-14.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\15307130191-16-17.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\15307130191-16-20.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130040-01-10.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130040-01-13.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130040-01-15.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130040-05-11.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130040-12-15.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130198-15-03.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130198-16-05.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130228-12-02.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130228-12-04.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/trainset\16307130322-12-08.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/valset\16307130343-12-19.wav
There aren't any elements to reflect in axis 0 of `array`
./tmpwav/valset\17307130106-10-20.wav
There aren't any elements to reflect in axis 0 of `array`
```

图5：对所有语音数据进行端点检测之后，在生成语谱图的过程中，会出现语音数据为空的情况，图中为具体信息。

3.2.3数据集清洗过程

对于产生语谱图质量过差的4个人的语音数据，采用直接剔除的方法。对于端点检测之后缺少的19个语音数据，采用从同一个人的同一类语音数据中选择一个用来补全的方法。

3.2.4所使用数据集分析

通过调查，原始语音数据集的大小为750 MB，而经过端点检测之后的语音数据

集的大小为407 MB，即端点检测的截取率为54.26%，这与我们录音过程中的经验性感受相同。我们在录音过程中往往在进度条走到中点之前就已经将孤立词说完。说明端点检测的准确率比较高。
另外，因为图片大小比较大，导致数据量比较大，所以在最后制作数据集的过程中，参考了CIFAR10数据集的制作方式，本项目将2400个图片作为一组，制作成一个数据集，最终训练集分成了4个小数据集，验证集分成了1个小数据集。这样可以避免numpy数组最后合并所需时间越来越长的问题。

4.实验结果分析与讨论

4.1最佳实验结果

本项目最终在验证集上的正确率为：86.95%。
各类别正确率见表2。

词	正确率/	词	正确率/	词	正确率/	词	正确率/
数字	75.00	总工	63.75	复旦	83.75	Process	86.25
语音	63.75	北京	82.50	饭店	88.75	Print	90.00
语言	86.25	背景	81.25	Speech	98.75	Open	82.50
识别	91.25	上海	98.75	Speake	95.00	Close	97.50
中国	93.75	商行	91.25	Signal	83.75	Project	93.75

表2：各类别正确率。

4.2分析与讨论

基于图像分类的模型在验证集上达到了较高的正确率。表明了本项目所制作的数据集具有较高的质量，即验证了本项目提出的端点检测算法的有效性。同时，
这个结果也表明语谱图作为语音的一个特征在孤立词识别任务中的有效性。最后本模型的结果证明了图片大小变换算法成功地解决了语音识别领域语音不定长问题，表明了这种在图像角度解决语音不定长问题的算法是行之有效的。

5.结论

本项目基于图像分类解决孤立词识别问题，达到了较高的正确率。本项目提出了一种独特的端点检测算法，基于帧和一个定长缓冲区，成功地解决了一个语音中有多段有声段的端点检测问题。本项目并没有从语音领域出发，而是从图像识别领域出发，以语谱图图片作为特征，较好地解决了孤立词识别问题，体现了自然语言处理和计算机视觉两大领域之间的联系和互相促进的作用。本项目提出了一种颇具创新性的解决语音不定长问题的算法，先绘制出语音的语谱图，然后在图像处理的角度将语谱图转化为固定大小，从而解决了语音不定长问题，为解决语音不定长问题提供了一个新的角度。

6.参考文献

[1] ADVANCING ACOUSTIC-TO-WORD CTC MODEL Jinyu Li, Guoli Ye, Amit Das, Rui Zhao, Yifan Gong

[2] Towards End-to-End Speech Recognition with Recurrent Neural Networks Alex Graves, Navdeep Jaitly

[3] Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks

Ying Zhang, Mohammad Pezeshki, Philemon Brakel, Saizheng Zhang, C ‘esar Laurent, Yoshua Bengio, Aaron Courville

[4] https://github.com/wiseman/py-webrtcvad

[5] https://librosa.github.io/librosa/

[6] https://pytorch.org/

致谢

在此感谢两个室友提供的帮助和想法。