编　　号：_____

审定成绩：_____

重庆邮电大学
毕业设计（论文）

中文题目	基于神经网络的语音情感识别算法的研究与实现
英文题目	**Research and Implementation of Speech Emotion Recognition Algorithm Based on Neural Networks**
学院名称	自动化学院
学生姓名	瞿荣辉
专　　业	物联网工程
班　　级	**08051403**
学　　号	**2014212724**
指导教师	姓名 谢昊飞　　职称 教授
答辩组负责人	姓名 王　平　　职称 教授

二零一八 年 五 月

重庆邮电大学教务处制

自动化 学院本科毕业设计(论文)诚信承诺书

本人郑重承诺：

我向学院呈交的论文《基于神经网络的语音情感识别算法的研究与实现》，是本人在指导教师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明并致谢。本人完全意识到本声明的法律结果由本人承担。

年级 2014

专业 物联网工程

班级 08051403

承诺人签名

2018 年 4 月 30 日

学位论文版权使用授权书

本人完全了解重庆邮电大学有权保留、使用学位论文纸质版和电子版的规定，即学校有权向国家有关部门或机构送交论文，允许论文被查阅和借阅等。本人授权重庆邮电大学可以公布本学位论文的全部或部分内容，可编入有关数据库或信息系统进行检索、分析或评价，可以采用影印、缩印、扫描或拷贝等复制手段保存、汇编本学位论文。

（注：保密的学位论文在解密后适用本授权书。）

学生签名： 指导老师签名：

日期： 年 月 日 日期： 年 月 日

摘要

随着人工智能技术的不断发展和突破，如何从被动交互转换为主动交互，将是智能产品必备要素，因此，评估用户的情感必然是人工智能的发展趋势。基于语音信号实现情感识别，可以很好地感知用户的情感变化，进而采取相应的主动交互请求。

本文构建了基于 AlexNet 框架，融合 DTPM 池化和 SVM 分类器的语音情感识别模型。首先，从原始的一维语音信号中，采用 matlab 实现 MFCC 二维对数梅尔谱图的提取，将其重组为三通道的对数梅尔谱图（static, delta and delta-delta）作为 AlexNet 框架的输入；其次，迭代训练 AlexNet 模型，微调每层结构中的 weight 和 bias；然后，将 AlexNet 框架中全连接层 Fc7 的输出的特征向量 n*4096-D，通过 DTPM 池化，在离散的语音特征中加入空间维度，将其组合成固定的 7*4096-D 为特征向量；最后，将固定长度的 7*4096-D 特征向量，通过 SVM 进行分类，得到七种情感的置信标签。

本文使用 EMO-DB 语料库作为数据集，进行 AlexNet 框架的迭代训练，得到最优权重和偏置。当只使用 AlexNet 模型测试置信度只达到 60%；结合 DTPM 池化和 SVM 分类器后进行测试，置信度达到了 83%；通过交叉验证方法提高模型的泛化能力，置信度下降至 74.7%（文献 18 置信度 87.31%）。最后，对自然录制语音进行测试，模型的最高置信度为 58%（文献 18 对自发 BAUM-1s 视听数据平均置信度 44.61%）。

考虑到模型的实际应用场景，将 AlexNet 网络部署到了安卓端，用户可以基于该应用，可以直接通过手机等智能终端录入语音并进行情感识别，此时得到最高执行度和对应情感标签。

关键词：语音情感识别，AlexNet，DCNN，DTPM，SVM

Abstract

The success of artificial intelligence technology has enabled intelligent products to transition passive interactive mode to active interactive mode. Therefore, how to evaluate the user's emotions will become the trend of artificial intelligence. This article based on the voice signal to achieve emotional recognition, a better perception of the user's emotional changes, and then take the appropriate active interactive requests.

This paper constructs a speech emotion recognition model based on AlexNet framework, DTPM pooling and SVM classifier. First, from the original one-dimensional speech signal, Matlab was used to extract the MFCC two-dimensional logarithmic mel spectrum, recombining it into a three-channel logarithmic melline spectrum (static, delta and delta-delta) as the AlexNet input ; Second, epoches trains the AlexNet model, fine-tune the weight and bias in each layer structure; Then, the feature vector n*4096-D of the output of the fully connected layer Fc7 is pooled by the DTPM, in discrete speech features added the spatial features and combined into a fixed 7*4096-D feature vector. Finally, the fixed-length 7*4096-D feature vectors are classified by SVM to obtain seven emotion confidence labels.

In this paper, EMO-DB corpus is used as a data set to perform iterative training of AlexNet framework to obtain optimal weights and biases. When only the AlexNet model is used, the test confidence is only 60%; when the test is combined with the DTPM pooling and the SVM classifier, the confidence level reaches 83%; the cross-validation method improves the generalization ability of the model, and the confidence level drops to 74.7% (Document 18 Confidence 87.31%). Finally, testing the nature of recorded speech, the highest confidence level of the model was 58% (the average confidence level of the audiovisual data of the spontaneous BAUM-1s in the literature 18 was 44.61%).

Taking into account the actual application scenario of the model, the AlexNet network is deployed to the Android client. Based on the application, the user can directly input voice and carry out emotion recognition through a smart phone such as a mobile phone. At this time, the highest confidence level and corresponding emotion tag are obtained.

Keywords: speech emotion recognition，AlexNet，DCNN，DTPM，SVM

目录

第 1 章 引言

1.1 研究背景和意义

随着大数据与计算能力的提升，人工智能技术在近十年的时间里发展迅速，深刻地影响着我们的工作、学习和生活。同时，人们对于这项技术的期待也逐渐提高，不仅停留在简单的被动交互层面，更多的是希望能与计算机主动交互。因此，评估用户的情感必然是人工智能的发展趋势。掌握用户的情感信息，一方面可以了解用户的业务需求、方便提供更准确的服务与帮助；另一方面，可以记录用户的生活信息，及时反馈用户的情感状态变化，帮助用户了解自己、也使得计算机变得更加智能。

语音信号是最自然地交流方式，不仅承载着明确的语言内容，也含有说话者的隐性语言信息。语音携带着的感情不同，给人的感觉也完全不一样。现如今，计算机在人们的日常生活中占据着不可替代的地位，人机交互也受到了广泛关注，建立更加自然而友好的人际交互系统。

过去的二十年里，人类在从语音信号中识别情感即语音情感识别方面付出了巨大的努力。根据文献[1]，“Emotional intelligence” 在人机交互中是必不可少的研究方向。

在人机交互系统中，计算机等智能终端如何能够体会到人的喜怒变化并见机行事呢？情感识别就是想要建立一个能够自动提取用户语音特征、识别情感类型，并能主动做出准确反应的系统，这样的系统可以像人一样分析用户的情感特征，理解用户的真实需求，并完成相应的任务。

情感识别在生活中有着非常广泛的应用前景，任何涉及主动交互和认知过程的系统中，都需要该技术的部署。

在情感计算方面，国内外的学者的研究主要集中在 voice library 的建立、extract feature、emotion recognize 等。要实现智能设备的主动交互，首先就必须让它懂得人类的情感讯息。语音是人类最普遍的交流方式，speech emotion recoginize 在人机情感交互中占据首要地位，情感识别的应用在智能交互，信息检索，语音通行，娱乐行业等：

（1）信息检索

现如今，搜索工具已不仅仅是文本检索方法，越来越多的学者将目光聚焦于多媒体检索方法，而情感识别将在多媒体检索技术中占据着不可替代的作用。

（2）电话服务

许多公司为了节约成本，在客服系统中采用计算机自动电话服务系统。但现在的自动电话系统过于机械化，语言表达没有感情，将语音情感识别技术加入电话服务系统中会使电话服务变得更为人性化，它能够识别客户的情感并作出相应的回答，使客户体验效果更好。

（3）娱乐行业

曾经在日本风靡一时的玩具“虚拟宠物”，是一种只能表达情感而不能识别人的情感的简单玩具，它具有鸡蛋大小的外形，只要点击三个按钮中的一个，就会作出相应的情绪相应。虽然这个宠物只能单向地表达情感，却能影响着人类的情感。如果这个宠物拥有识别人类情感的功能，无疑会更加丰富人类的生活。

以上都是语音情感识别技术在实际生活和生产中应用的领域，当然，也不仅仅局限于这些领域，在刑事安全，计算机辅助教学等方面，都可以应用语音情感识别技术。

系统实现的前提便是语音信号特征的提取，特征提取是弥补语音信号与主观情绪之间情感差距的关键步骤。到目前为止，手工设计提取的各种特征已被用于语音情感识别。然而，手工设计所提取的特征通常是低级的，不足以描绘出主观情绪的全部特征，自动学习特征提取的研究，将是主要方向。

到目前为止，由于在人机主动交互系统的广阔前景，Speech Emotion 可以更好的体现人们主观意识的情感，在面部识别不能很好的捕捉的时候，语音就能更好的体现出最直接的情感。

1.2 国内外研究现状

1.2.1 国外研究现状

早期的学者们论证了对 Emotion lable 进行定义和要求。文献[2]中的一个 Idea 表明：“Computer has emotions”[2]，吸引了众多的科研机构和研究学者，纷纷踏上探索语音情感识别的道路，把人工智能和语音情感融合起来，成为了众多学者追求的目标。

在这段时间里，美国麻省理工学院首先取得了突破，对用户的各种情感信号进行了采集，比如：voice signal、Face image signal、Various physiological indicators，并能基本识别出用户的各种情感，还能简单的回应[3]

90 年代末期，有学者提出了一个线性模型来衡量语音和情感之间的联系。该模型应用在了电子商务系统中，虽然准确率不高，但也表征着语音情感在实际场景中的初步应用。

总的来说，情感识别在这段时间仍然处于初级起步阶段，虽然很多有价值的研究成果陆续发表，但是没有形成一套广泛认可、系统、合理的理论和研究方法。

进入 21 世纪后，Speech Emotion recognize 在人工智能设备的普及之下，将会取得更大的发展 [4]。

其中比较出名的有：2005 年，首次举办 Affective Computing and Intelligent Interaction 会议；2009 年，首次创立了语音情感类的学术竞赛；2010 年，创办《IEEE 情感交易计算》期刊；2011 年，举办首届国际视听情感挑战与研讨会(AVEC)学术竞赛[4]。

经过 10 余年的探索和研究，在情感模型框架的搭建、情感语音数据库的构建、语音情感特征的分析、语音情感的分类等领域取得了很好的发展[2]。

1.2.2 国内研究现状

国内外学者，都致力于从语音信号中识别人类情感。因此，如何对语音信号表征进行提取、训练，俨然成为了国内外共同的研究目标。

清华大学提出“Interpersonal harmony”的概念，实现了 Interpersonal harmony 与 media 的集成，达到了人机和谐交互的目标。

中科院通过基频曲线生成模型，对未进行处理的语音进行模拟得到对应的语音曲线。让合成效果的语音，表征更加自然和口语化[5]。

东南大学提取语音的 84 个特征参数，包括短时能量、基音、发音和不发音帧数、共振峰等。使用线性判别分析进行情感特征降维和 fisher 判别准则，通过高斯混合模型识别出五种情感类别[6]。

华中科技大学的语音识别与合成，建立了语音云。减弱手工 feature 与 semantic content 的“semantic gap”。验证了 32 维的特征向量在识别效果上，远高于目前广泛的 21 维的情感向量。所以，人工神经网络模型的改进和提高将更好的提高语音情感识别的准确度[7]。

浙大的人工智能研究所和中国科院的语言研究所等。对不同语音特征的 speech library，从低维度到高维度，使用 Different dimensionality reduction algorithms，提高模型的整体泛化能力。算法更好的提取出语音特征，去除外部非语音信号，体现出更优的语音识别效果[8]。

结合国内外的发展现状，语音情感识别一直都收全球各位学者的关注，无论是国外的研究团队，还是国内的研究学者。他们都在运用更好的方式提取语音的特征表示，以减小语音特征和语音情感内容之间的 feature map ，从而实现语音情感识别的最优效果。因此，本文使用 DCNN 提取语音特征，运用 AlexNet 框架搭建语音情感识别系统是有意义的。

1.3 主要内容和工作安排

通过对国内外研究现状的调研，发现语音情感识别主要的难题是如何选择语音信号的特征、如何解决语音信号的随机性和多变性、如何选择合适的算法进行情感分类以及最终的实际应用部署问题。因此，本文主要研究内容如下：

首先，基于 MATLAB 对 wav 语音进行预处理、傅里叶变换和倒谱分析，提取语音信号的梅尔频率倒谱特征 MFCC；将 wav 文件转化为基于谱的 RGB 图片，作为 AlexNet 框架的输入数据；

其次，重点讲述卷积神经网络算法的 AlexNet 框架，基于 EMO-DB 语音库，对 AlexNet 框架 Training 过程中的输入输出、卷积池化进行讲述；在 AlexNet 模型的基础上融合 DTPM 池化和 SVM classification，构建出本文的语音情感识别模型。

然后，进行 AlexNet 模型进行迭代训练，微调框架中每一层（8 层）的参数 weight、bias，输出最优模型的 pb 参数文件，方便模型的移植；Tensorboard 可视化训练过程，得到训练和测试结果曲线；接下来，将 AlexNet 模型训练过程中的 Fc7 输出 n*4096-D 特征向量通过 DTPM 池化，在原始特征向量中加入空间维度，fixed 7*4096-D feature，作为 SVM 的 feature_input，从何提高模型情感测试的置信度，输出语音情感测试的准确度。

最后，实现模型在 Android 端的部署，在 Android 端加载 opencv 库进行语音特征 MFCC 的提取；加载 TensorFlow 平台，调用模型训练得到参数.pb 文件；进行自然条件下录制语音的情感识别。

论文的整体框架如下图 1.1 所示：

图 1.1 论文架构

第 2 章 语音情感识别的基础理论

本章通过介绍基本的语音情感相关知识，和使用的网络架构，从而展现出论文的知识体系结构如图 2.1。

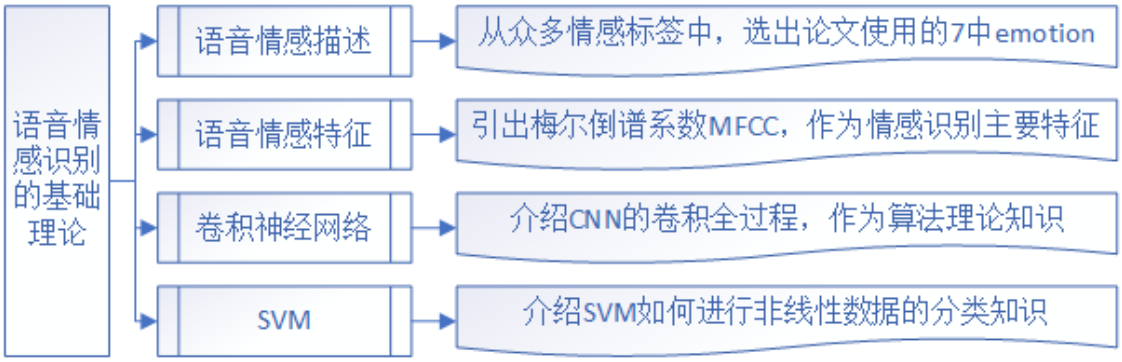


图 2. 1 语音情感识别的基础理论

2.1 语音情感描述

如何对人类表达的情感进行分类和定义，是研究语音情感识别的最基本的问题。目前国内外并没有制定一个统一的标准去衡量语音情感。

在早期的情感相关研究中普遍使用这种模型。人类描述情感的词语丰富多样，如何找到更为普遍，更具有研究价值的情感分类是语音情感识别的基本问题。

学者	基本情感
Arnold	Anger,aversion,courage,dejection,desire,despair,dear,hate, hope,love,sadness
Ekman,Friesen,Ellsworth	Anger,disgust,fear,joy,sadness,surprise
Fridja	Desire,happiness,interest,surprise,wonder,sorrow
Gray	Desire,happiness,interest,surprise,wonder,sorrow
Izard	Anger,contempt,disgust,distress,fear,guilt,interest,joy, shame,surprise
McDougall	Fear,disgust,elation,fear,subjection,tender-emmotion, wonder
Oatley,Johnson-Laird	Anger,disgust,anxiety,happiness,sadness
Panksepp	Anger,disgust,anxiety,happiness,sadness
Emotion[9]	Anger,fear,surprise,sadness,disgust,happness,nertual

图 2. 2 基本情感的类别定义

2.2 语音情感特征

广泛用于情感识别的情感语音特征大致分为四类：1）声学特征; 2）语言特征，如词汇信息; 3）语境信息，如主题，性别，文化影响，4）混合特征，如上述两个或三个特征的整合。

声学特征主要包含韵律特征，语音质量特征和光谱特征。音高，响度和持续时间代表韵律特征，因为它们表达语言的压力和语调模式。语音质量特征作为单个语音的特征听觉色彩，在表达积极或消极情绪时显示出有区别性

韵律特征和语音质量特征相结合表现出比单独使用韵律特征更好的性能[43]。近年来，声门特征和语音源参数已被用作语音情感识别的更高级语音质量特征。根据声音的短期功率谱 LPCC，LFPC 和 MFCC 计算频谱特征。近年来，还研究了基于局部 Hu 矩的来自听觉启发的长期谱时间表示的调制谱特征[10]和基于局部 Hu 矩的加权谱特征[11]。

2.3 卷积神经网络

2.3.1 CNN 的结构

典型的卷积神经网络由多层神经元组成，每层根据操作的不同，通常可以划分为 Input_layer、Output_layer、Pooling_layer、FC_layer and Output_layer，一层一层串行地提取图像特征。例如比较常用的 CNN 网络——LeNet-5，使用的卷积神经网络的结构如下图 2.2 所示，包含 7 层网络结构[12]。

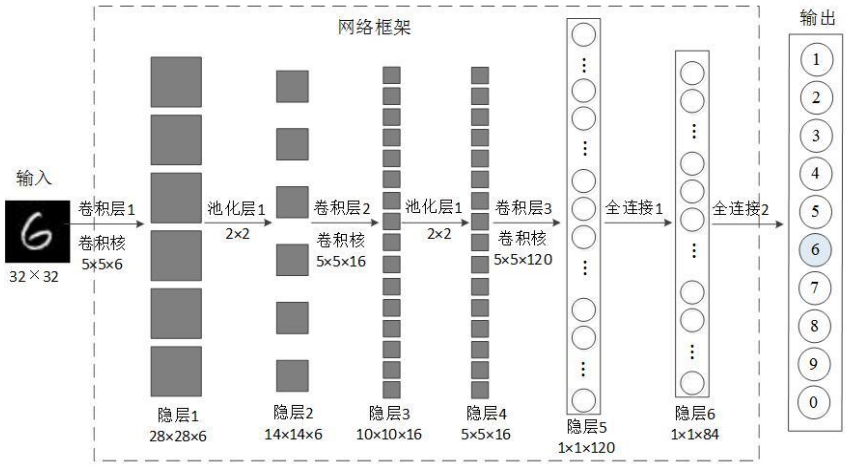


图 2. 3 CNN 基本结构

输入层加载原始的图片，原始图片大小是 32*32 像素。卷积层 1（C1）对应有 6 个卷积核，每个 convolution kernel _size = 5*5。 Output_layer product 6 feature map，每个特征图大是（length – kernel_size+1）*（width-kernel_size+1）=28*28。此外，每个 filter_size = 5*5，1 个 bias，特征图内部共享 same_filter，这就大大减少了参数数量。输入层与 C1 层中连接的参数数量总共有（5*5+1）*6=156 个。然后再送入激活函数中，即可得到 C1 层的特征图。

接下来，C1 层对原始图片特征提取后，会经过一个池化采样的操作，这个步骤的主要作用是通过降低图片的分辨率来压缩图片、减少参数数量，比例缩放的形变并不会改变图片的空间特征，这样就得到了池化层 2（S2）层池化层的图片，这个过程也可以叫做下采样层的操作。这一层的图片大小为 14*14 像素，因为 C1 层中的 28*28 像素的图片经过一个 2*2 的池化，图片的长度和宽度分别缩小 2 倍，得到（28/2）*（28/2）=14*14 的图片，特征图个数不变，仍然为 6。

S2 层再经过一个卷积操作后得到卷积层 3（C3）层的图片，卷积核仍然是 5*5，得到的 C3 层图片大小为（14-5+1）*（14-5+1）=10*10。接下来，C3 层仍然要进行池化或者说下采样操作，较少神经元的个数从而减少参数数量来提高网络学习过程中的训练效率。与 S2 层一样，池化层 4（S4）层仍然采用 2*2 的池化方式，最终得到的特征图大小为（10/2）*（10/2）=5*5 的像素，特征图个数仍然保持不变，为 16 个。

卷积层 5（C5）层仍然是 S4 层经过卷积获得，卷积滤波器核依然为 5*5，由于 C5 层中的特征图为 5*5 的像素，那再经过 5*5 的卷积核后，得到的特征图就变成（5-5+1）*（5-5+1）=1*1。这两层的相连方式中，S4 层中的 16 个特征图的所有神经元都与 C5 层中的 120 个单神经元全部相连，所以这两层的连接方式也可以看做是全连接的方式。

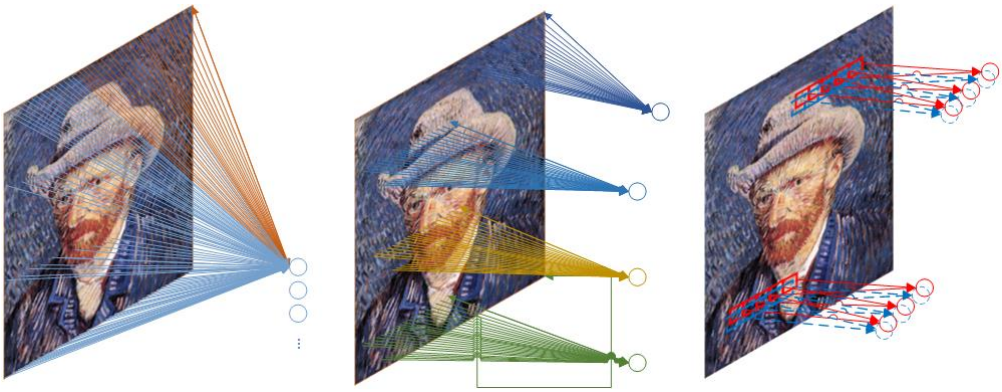
全连接层 6（F6）层与 C5 层采用全连接的方式，F6 层 Input_cell = 120，Output_cell = 84，all_output_layer = Input_cell * Output_cell + Input_cell =10164。

FC_7 Input_data = 84，Output_data = 10，数据集中最后的输出分类为 0-9，那么对应的 Output_layer = 10。

F6 层 Gaussian Connection[12]的负似然数值连接到每个径向基函数，那么当给定一个输入模式时，损失函数应当使得欧式径向基函数 RBF 单元与期望的输出模式更加贴近。

2.3.2 CNN 的卷积

建立一个输入为 500x500 像素图像的人工神经网络，Input_data = 500x500，hide_layer = 10^6 cell。下面比较全连接的 ANN 和部分连接 ANN 权重参数的数量[13]。



(a) 全连接　　(b) 局部连接　　(c)卷积

图 2. 4 CNN 卷积过程

当建立全连接的 ANN 处理图像时，输入层和隐藏层之间的权重需要 500*500*10^6=25*10^{10} 个，如图 2-3（a）所示。使用全连接的 ANN 用于大图像的处理时面临着参数过多、计算量无法承受的问题。

参考人眼在接受外部图像信息的过程，即每次只看到部分图像的信息，通过每次感受的局部图像理解完整的图像。这就是使用滤波器实现输入层和隐藏层之间的局部连接。设置 10*10 的滤波器模仿人眼感受局部的图像区域，这时，一个隐藏层的神经元通过滤波器和输入层中一个 10*10 区域连接。隐藏层有 10^6 个隐藏神经元，因此隐藏层和输入层之间就有 10^6 个 filter，相应的权重数是 10^8 个，如图 2-3（b）所示。

全局连接改成局部连接已经将权重数量从 25*10^{10} 个减少到 10^8 个，但是权重过多计算量巨大的问题依然没有解决。经过研究发现图像有一个固有的特性——一部分统计特征与其他部分是一样的，也就意味着从某一局部学习到的特征也可以用在另外一个局部学习特征。所以，可以采用同样的特征矩阵去学习一张图象

的所有特征。在卷积神经网络中，一个滤波器就是一个权重矩阵，通过该矩阵去提取图像所有的特征信息。

同一个滤波器自然就可以应用于图像中的任何地方，10^6 个 10*10 滤波器对应的是图像中 10^6 个不同的 10*10 区域。如果这些滤波器完全相同就是将局部特征应用于整个的图像，如图 2-3（c）所示。所有的滤波器 shared weights，这样 Weight_size 从 $25*10^{10}$ decrease 100 个左右，大大减少权重参数的数量及计算量。

这里使用的局部连接和权重共享的概念就是实现了卷积操作，10*10 的滤波器就是一个卷积核。一个卷积核表示图像的一种局部特征，当需要同时表示图像的多种局部特征时可以设置多个卷积核。图 2-3 中第一个卷积层使用了 6 个卷积核，通过卷积操作得到 6 个隐藏层的特征图。

2.3.3 CNN 的 Pooling

通过卷积获得了特征图或者说图像的特征，但是直接使用卷积提取到的特征去训练分类器，仍然会面临非常巨大的计算量的问题。比如：Input_data 是 96*96，400 个 fliter，每个 filter_size 为 8*8，那么经过卷积操作后，得到的图像维度为 (96-8+1)*(96-8+1)=89*89=7921，由于有 400 个卷积核，所以得到 400 个特征图，那么得到的总的图像维度为 $89^2*400=3,168,400$，这个计算量还是很大[14]。

根据图像的“静态性”属性特征，可以对图像特征进行聚合统计操作，例如可以对图像的某一部分进行平均值操作或者取最大值，得到低维度的新图像。在卷积神经网络中，这种对特征进行降维的操作就是池化。池化的方式有 average_pooling /max pooling 和 L1-池化。图 2-4 是一个 max_pooling 的示意图，输入图像为 6*6。

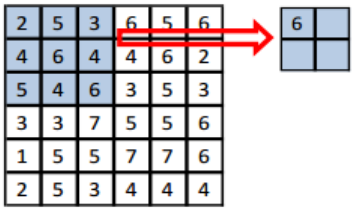


图 2. 5 CNN 的 max_pooling

进行一个 3*3 的最大值池化，即选取 3*3 邻域中的最大值 6，最后得到的池化结果是像素为 2*2 的矩阵。

2.4 SVM 支持向量机

SVM 的主要思想是将数据集描述为空间中的离散点，找到一个超平面实现对于所有分类的划分，容易实现二分类问题，对于多分类任务则进行递归操作，直到完成所有类别的分类。如果线性不可分时，则通过非线性映射的方式将数据集投影到高维空间中，基于更高维度的超平面实现分类[15]。

2.4.1 SVM 原理

假设已经获得数据集，SVM 就是要找到一个超平面将不同类别的数据划分为不同的区域，当然在空间中不止一个平面可以满足这个要求，如图，所以，我们需要找到一个距离各类样本最小的平面，才能尽量提高准确率。

超平面的线性划分如图 2.5 所示：

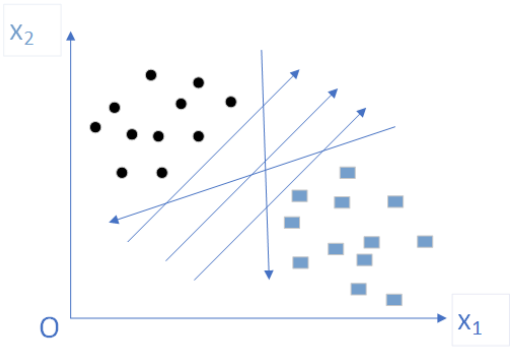


图 2. 6 线性超平面

（a）空间中的平面方程：

$$Ax+By+Cz+D=0 \tag{2.1}$$

转化为矩阵方程：

$$\begin{aligned} &w_1x_1+w_2x_2+\cdots+w_nx_n+b=0(n=1,2,\cdots n)\\ &w^Tx+b=0(T\text{表示转置}) \end{aligned} \tag{2.2}$$

数据集中任意一点到超平面的距离为：

$$\gamma=(w^Tx+b)/\|w\|=f(x)/\|w\| \tag{2.3}$$

那么，当 $y(w^Tx+b)=1$ 时，支持向量刚好在边界上；而对于所有不是支持向量的点，则 $y(w^Tx+b)>1$，所以其正确分类的条件是：

$$s.t.\quad y_i(w^T x_i + b) \geq 1,\quad i = 1,2,\ldots,m \tag{2.4}$$

显然，SVM 的工作便可以转化为，找到最优的超平面，即使得 γ 最大时的参数 w,b 。

$$\begin{aligned} &\max_{w,b} \frac{2}{\|w\|} \\ &s.t.\quad y_i(w^T x_i + b) \geq 1,\quad i = 1,2,\ldots,m \end{aligned} \tag{2.5}$$

（c）最大化支持向量间的距离，只需求得最大 $\|w\|^{-1}\|\omega\|$，对分子分母求倒，等价于最小化 $\|w\|^2$，上述问题可等价转换为：

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &s.t.\quad y_i(w^T x_i + b) \geq 1,\quad i = 1,2,\ldots,m \end{aligned} \tag{2.6}$$

综上，SVM 模型可以转换为一个凸二次规划问题，考虑采用拉格朗日乘子法转换该优化问题，将目标函数和条件组合成一个方程式求解。

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(w^T x_i + b)) \tag{2.7}$$

对 $L(w,b,\alpha)$ 方程求 w,α 的偏导为零得：

$$\begin{aligned} w &= \sum_{i=1}^{m} \alpha_i y_i x_i \\ 0 &= \sum_{i=1}^{m} \alpha_i y_i \end{aligned} \tag{2.8}$$

于是得到上述问题的对偶问题：

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &s.t.\quad \sum_{i=1}^{m} \alpha_i y_i = 0,\quad \alpha_i \geq 0, i = 1,2,\ldots,m \end{aligned} \tag{2.9}$$

对上式进行求解，先求解得到 α ，然后求出 w,b ，即可得到模型：

$$f(x) = w^T x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b \tag{2.10}$$

2.4.2 非线性分类

为了寻找非线性数据的超平面，通过引入核函数 $K(.,.)$，将数据从低维映射到高维，从未找到最优超平面，实现非线性数据的分类问题。

对应的函数方程：

$$f(x)=\sum_{i=1}^{N} w_i\phi_i(x)+b, \tag{2.11}$$

ϕ 为数据从输入空间到摸个特征空间的映射，加入数据的对偶特性，将训练数据和测试数据点内积表示：

$$f(x)=\sum_{i=1}^{N} w_i y_i \langle \phi(x_i)\bullet\phi(x)\rangle + b, \tag{2.12}$$

kernel function：在 low-dimensional space 计算内积 $\langle \phi(x_i)\bullet\phi(x)\rangle$：

$$K(x,z)=\langle \phi(x)\bullet\phi(z)\rangle \tag{2.13}$$

那么，对于两个向量 $x_1=(\eta_1,\eta_2)^T, and, x_2=(\xi_1,\xi_2)^T$，进行内积后便为五维空间映射：

$$\langle \phi(x)\bullet\phi(z)\rangle=\eta_1\xi_1+\eta_1^{\,2}\xi_1^{\,2}+\eta_2\xi_2+\eta_2^{\,2}\xi_2^{\,2}+\eta_1\eta_2\xi_1\xi_2 \tag{2.14}$$

如果当 $\langle \phi(x)\bullet\phi(z)\rangle^{n}$ 是 n 维的内积的时候，这样的计算量是很大的，因此直接在低维空间的计算，我们用核函数能简化表示。分类函数可以表示为：

$$f(x)=\sum_{i=1}^{n} \alpha_i y_i k(x_i,x)+b \tag{2.15}$$

In the linear unseparable function: 将数据映射到 higher space，加入 Lagrange 因子 α，将 constraint condition 融合到 objective function，求一个式子的最值；通过引入核函数 $K(x,z)=\langle \phi(x)\bullet\phi(z)\rangle$，将 n 维的内积在低维空间中进行计算。不断调节函数 $f(x)=w^Tx+b$ 中的参数（w,b），使得平面两边的数据到超平面的距离之和最小，完成 svm 模型的训练。

第 3 章 语音情感特征提取

梅尔到频谱系数 MFCC，能有效反映语音基于谱的特征，常用于语音识别的特征向量，对人类听觉系统进行建模 [16]，本章讲述 MFCC 特征提取的流程。

本章通过 Matlab 实现 EMO-DB 语料库的语音特征 MFCC 提取，通过 matlab 自带的函数 plot 进行图形界面的展示。

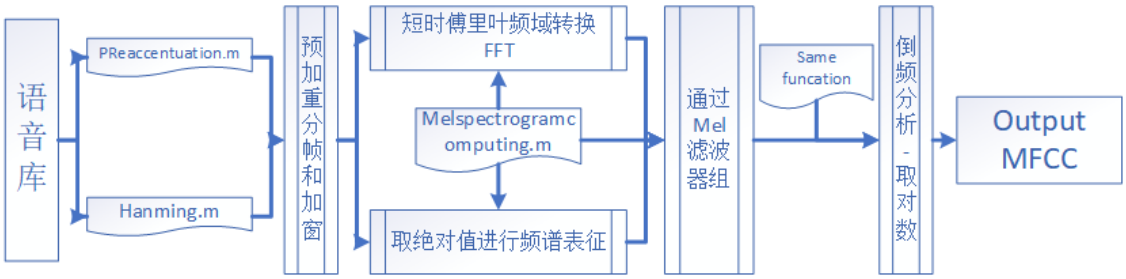


图 3. 1　MFCC 参数提取基本流程

3.1 EMO-DB 语料库

本文采用离散型的 EMO-DB 语音库[18]作为训练数据库，数据库由 10 位专业的讲德语的演员（5 名女性和 5 名男性）模拟这些情绪，由 10 条能够用于日常交流的德语发音组成（5 个短和 5 个更长的句子）。语料库包含 535 种情感语音，和 7 种情感标签。

3.2 梅尔频谱系数

语音信号的产生由声道的 shape 决定，声道的 shape 在语音短时功率谱的包络中显示出来。包络特征可以通过 MFCCs 准确描述。

Mel(f)与 f 的关系如下所示：

$$Mel(f) = 2595 \times \lg(1 + f / 700) \quad (3.1)$$

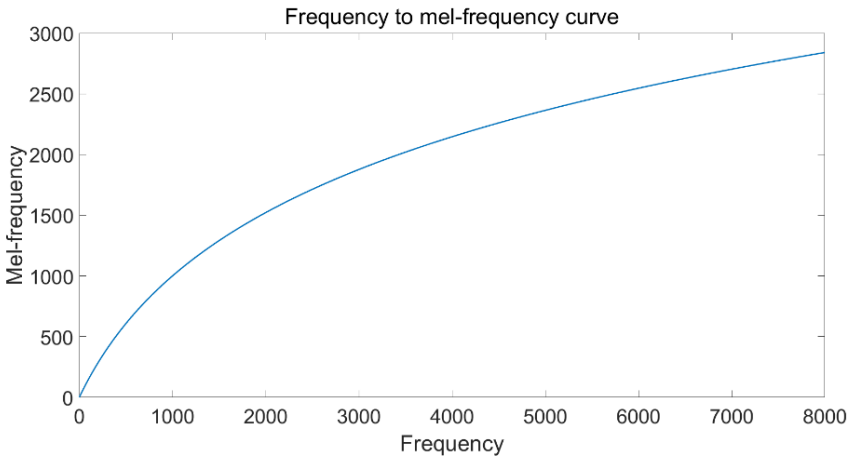


图 3. 2 Mel 频率与线性频率的对应关系

3.3 数据预处理

语音特征的提取，可以通过 python 和 matlab 进行提取，此处我们选择 matlab 进行提取，方便语音的可视化。在进行语音情感特征提取之前，与语音进行预处理可以提高语音质量，提升预测的准确度。

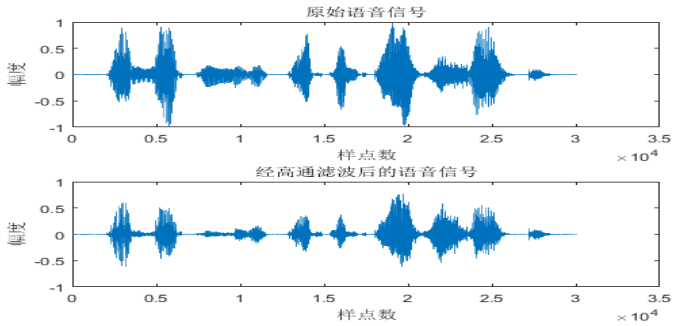
3.3.1 预加重

预加重的过程是高通滤波器处理语音的过程：

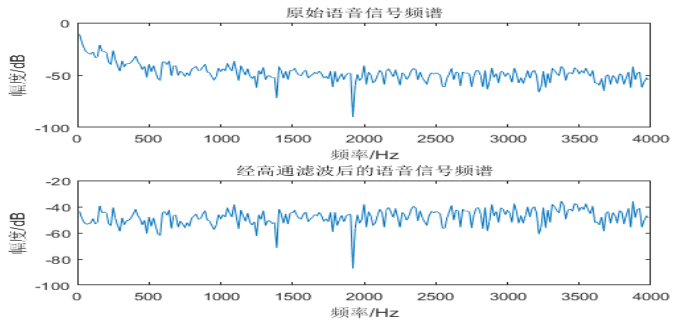
$$H(z) = 1 - \mu z^{-1}, \quad \mu = 0.97 \in (0.9 - 1.0) \quad (3.2)$$

语音信号输入时，低频成分能量集中，较高频成分更容易辨识。预加重的作用便是增强高频，削弱低频，让整个频域部分的波形变得平坦。

下图我们对 EMO-DB 语音库中的一个段级语音进行处理，通过 yujiazhong.m 模块对 03a01Fa.wav 进行遇加重处理，函数 plot 进行绘图显示。



（a）滤波前后的语音



（b）滤波后的语谱图

图 3. 3 预加重

3.3.2 分帧加窗

（1）分帧：

人耳对于语音的频段范围是 300-3400Hz，sampling frequency = 16KHz。文献中表明，信号只有超过 20ms 才可以有效的提取出情感信息，所以一般设置语音分帧长度为 20-30ms。通常情况下，为了保证分帧信号的连续性和平整性，一般会设置一段重叠区 10-15ms。

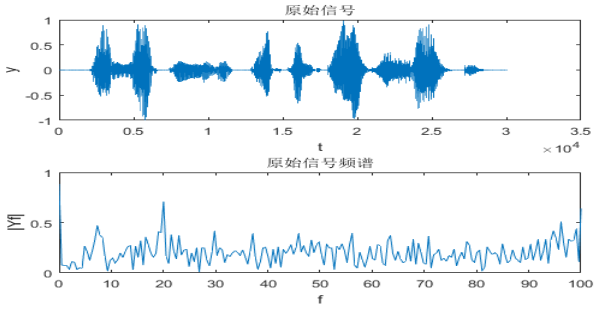
（2）加窗（Hamming Window）

如果将语音信号进行直接截断，即通过矩形窗进行处理，那么容易造成信号频域泄露的问题，为了减弱 frequency spectrum 泄露，通过 Hamming window 进行 framing。

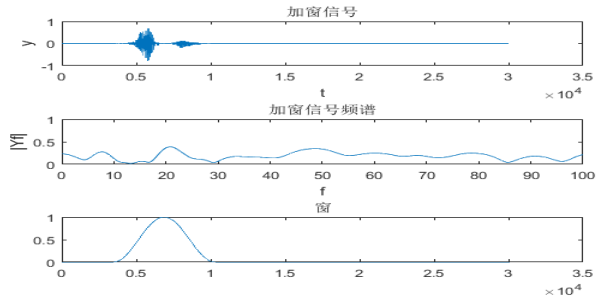
$$S'(n) = S(n) \times W(n)$$
$$W(n,a) = (1-a) - a \times \cos[2\pi n / (N-1)], 0 \le n \le N-1 \quad (3.3)$$

其中：a = 0.46，product different Hanmming window.

通过 Hanming.mm 模块对 03a01Fa.wav 语音处理，sampling rate = 16kHz，Hamming_size = 25ms，每次 Sliding Window = 10ms。下面将进行数据的 framing and window，得到了如下的 spectrogram：



（a）原始信号频谱图



（c）加窗后频谱图

图 3. 4 语音分帧效果放大图（假设窗口为 600ms）

3.4 快速傅里叶变换

Fourier transform 将 Time feature 转换为 Frequency feature，即表示为若干个谐波分量进行组合而成。通过观察频谱就可以获取到相关的特征信息。由于计算机处理的信号一般为离散信号，所以常采用离散傅里叶变换来对进行频域转换，为提升时间复杂度，进行 Discrete Cosine Transformation：

$$X_a(k) = \sum_{n=0}^{N-1} x(n) e^{-j^{2\pi k}/N}, 0 \le k \le N \qquad (3.4)$$

式中 x(n) 为输入的语音信号，N 表示 FFT 离散点的个数。通过 melspectrogramcomputing.m 模块对语音库进行 FFT 变换，下图是进过 FFT 计算得到的频谱图。

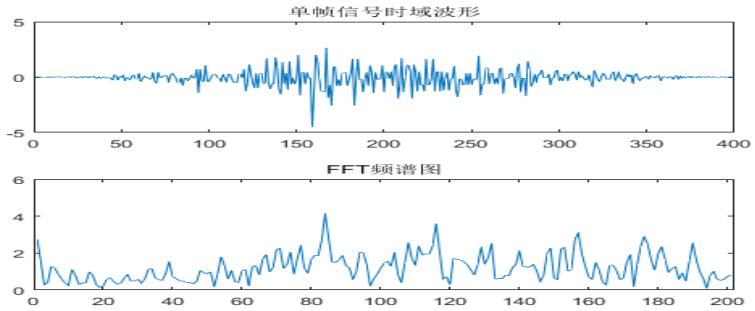
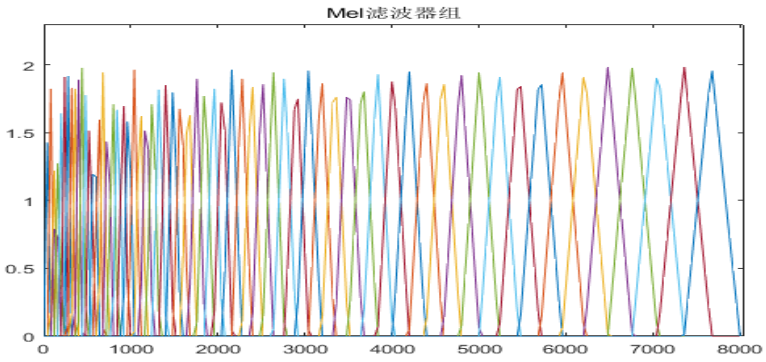


图 3. 5　FFT 频谱图

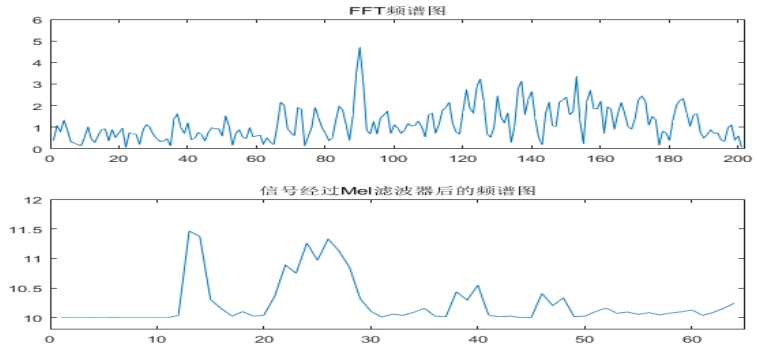
3.5 梅尔滤波器组

梅尔滤波器组是多个带通滤波器组合而成的滤波器组，多采用三角带通滤波器，Filter_size = 22-26。语音在 low-frequency，带通滤波器的间隔越小、幅值越大，将信号的能量谱输入就能很好地模拟人耳接收语音信号的处理过程。

同样，通过 melspectrogramcomputing.m 模块，将 FFT 的频谱图进过 Mel 滤波器，更好的模拟耳朵接受语音的过程。梅尔滤波器组频域波形如下图所示：



（a）Mel 滤波器



(b)通过 Mel 滤波器频谱图

图 3. 6 Mel_Fliter 后的平谱图

3.6 对数及 DCT 变换

人类的发声系统发出的语音信号是由基音信息与声道信息卷积而成。记作"s 卷积 v"。经过快速傅里叶变换后，卷积变成了乘法。即"FFT(s)*FFT(v)"。取对数后，乘变为"Log(FFT(s))+Log(FFT(v))"，卷积信号就转换成加性信号：

$$s(m)=\ln(\sum_{k=0}^{N-1}|X_a(k)|^2 H_m(k)),0\le m\le M \qquad (3.5)$$

DCT 变换得到语音信号的包络图，求出 L 阶的 MFCC 参数。

$$C(n)=\sum_{m=0}^{N-1}s(m)\cos[\pi n(m-0.5)/M],n=1,2,\ldots,L \qquad (3.6)$$

函数 melspectrogramcomputing.m，将 Mel 滤波器得到的频谱图再一次取对数，进行到谱分析，得到最终的 MFCC 系数，其波形图如下：

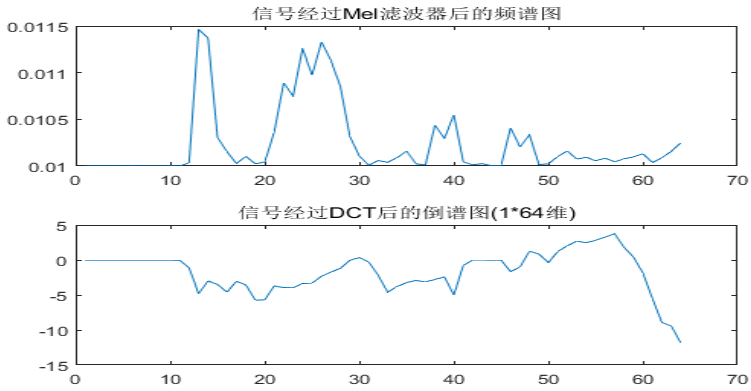


图 3.7 DTC 倒谱图，MFCC 系数

3.7 倒谱分析

对所得数据在进行一次 DCT 倒谱变换，得到梅尔倒频谱系数 MFCC。这是信号频谱的“高低频信息”，转换为倒谱上就对应于横轴上的不同取值，由于语音信号取决于频谱的“低频成分”，所以一般取 MFCC 的低阶系数即可。经过 DCT 后的倒谱图如图 3.8 所示。

对柏林语音情感数据库的每一段语音进行 MFCC 特征提取，并求取相应一阶差分、二阶差分。至此，每一帧已获得 64*3 维特征，再将 64 帧合并在一起，即得到 64*64*3 的数组。将该数组转换为 RGB 图片，即可输入到 CNN 网络中进行

训练。将其保存在相应文件夹中。结果如下图 3.9 所示，每段语音对应 n 张特征图（n=3,4,…）

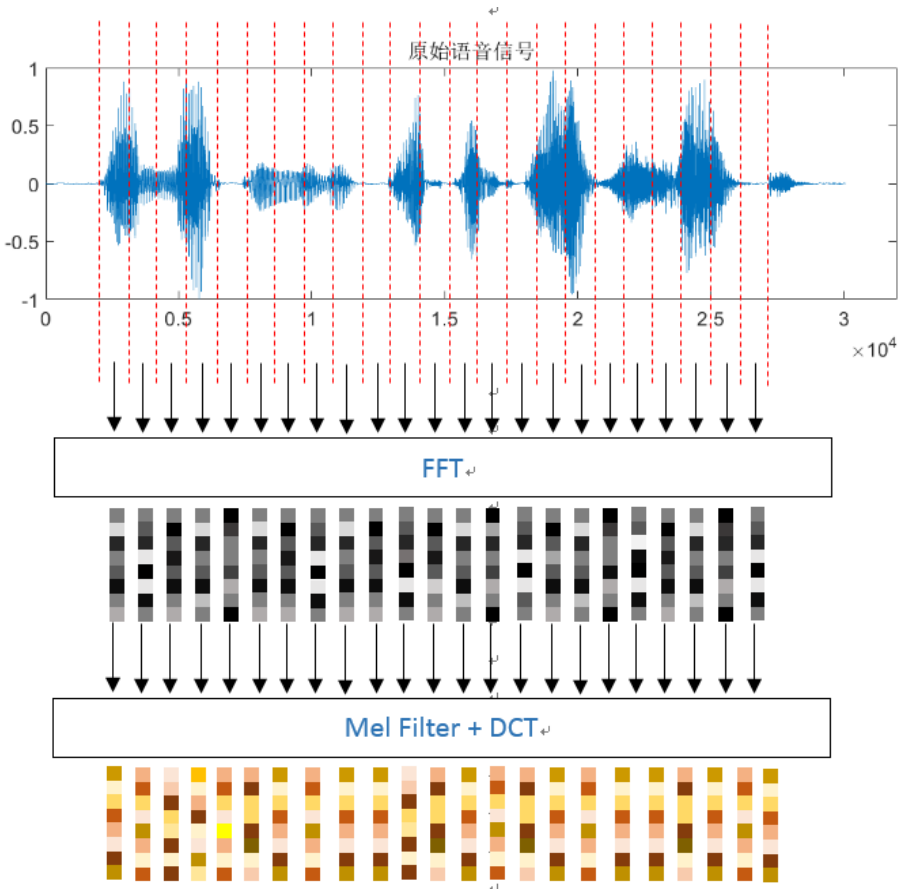


图 3. 8 DTC 倒谱图

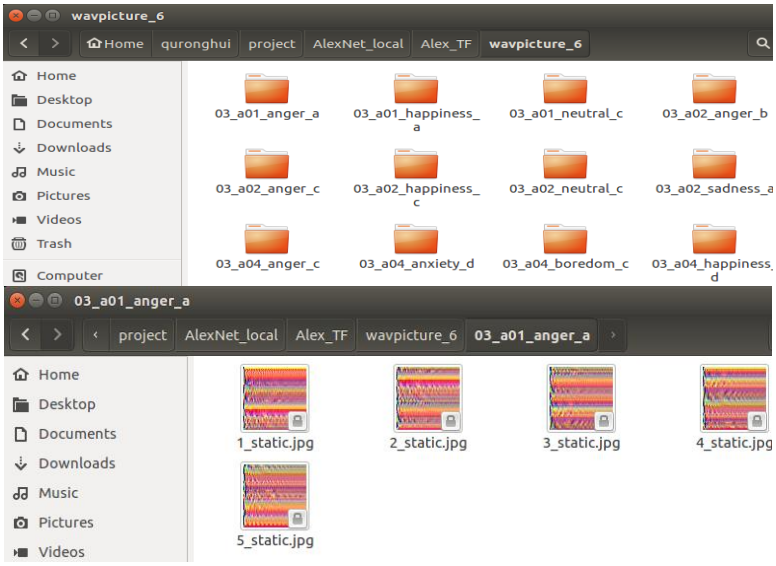


图 3. 9 每段语音对应的 MFCC 语谱图

第 4 章 基于 CNN 的语音情感识别模型

本章讲述 AlexNet 对语音情感特征 MFCC 的卷积过程，将全连接层 Fc7 输出的 n*4096 维特征向量；作为 DTPM 的输入进行时间池化，加上语音的空间维度，固定每段语音特征向量为 7*4096 维；最后通过 SVM 对维度数据进行情感标签的分类，得到语音情感识别模型。

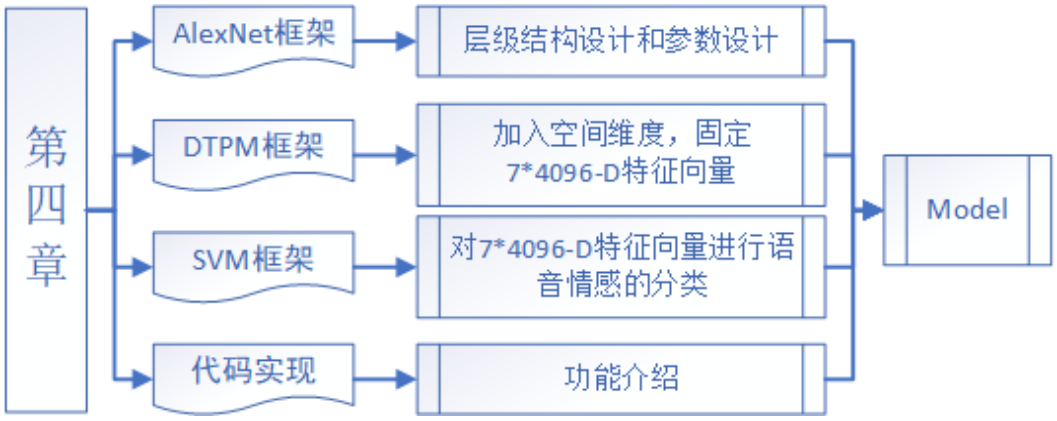


图 4. 1 语音情感识别模型

4.1 模型总体框架

论文《Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching》[18]中首次采用了 AlexNet 与判别时态金字塔匹配的方法实现了语音情感识别，给本文提供了参考。

原文中使用的是 AlexNet 和 DTPM 进行特征的训练，后续对数据进行 Lp 池化。本文的改进之处在于舍弃 Lp 池化，原因是 Lp 池化过程中 average 和 max 的状态不定，并且实现 Lp 池化需要增加训练难度和时间。因此，本文，将特征向量通过 SVM 分类器进行非线性分类，从而得到语音情感分类标签。

在本文中，我们使用 DCNN 学习语音信号的深度特征。如图所示，三个通道的对数梅尔谱图（static, delta and delta-delta）做为 DCNN 输入；DCNN 模型经过培训，可为每个细分群体提供深层功能；然后是线性 SVM 情感分类器；得到语音识别的基本框架，然后在德国情绪语音柏林数据集[17]（EMO-DB）进行训练。

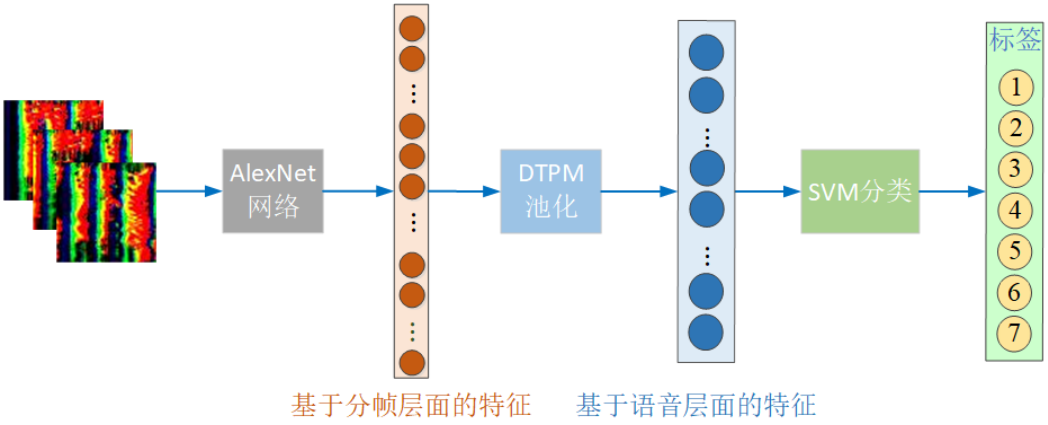


图 4. 2 模型总体框架

4.2 AlexNet 框架结构

AlexNet 框架是一个八层的网络结构，其良好的识别性能在 Image 中已经体现实现，因此论文将 AlexNet 迁移到语音识别中。它的结构包括五层卷积层、三层池化和三层全连接。论文[19]中已经说明，减少任何一个卷积层结果都变得很差，下面具体讲解每一层的构成。

4.2.1 函数定义

A. Dropout

Dropout 层通常用于降低特征维度，通过概率性丢失神经元进行。卷积过程中在 FC 全连接层后使用，避免过度训练形成的过拟合现象。Dropout 层设置选择概率大小为 p，当神经元通过该层后，输出的数据为原先的（1-p）倍，dropout 的训练过程如图 4-3 所示。

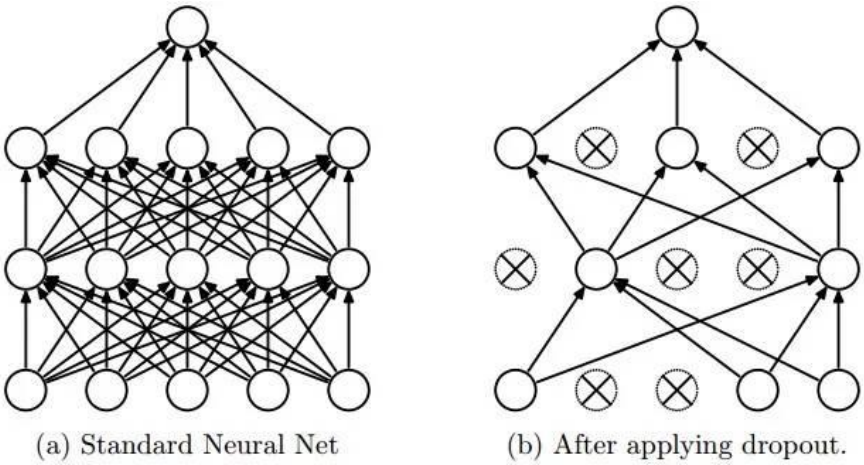


图 4. 3 dropout 神经元选择示意图

Dropout 进行神经元的选择丢弃时，擦用的是随机的形式，这样可以有效避免神经元的共适应（co-adaptation），产生新的语音特征。

B. Pooling

Alexnet 框架训练使用 max_pooling，相较于 CNN 使用的 average_pooling，AlexNet 让池化效果更明显。并且 AlexNet 中 stride 小于 kernel_size，增强了数据间的连接性，体现出更多的语音特征，有效提高识别能力。

C. LRN

AlexNet 使用在每一层卷积后对输出数据进行 LRN 局部归一化，Local normalization 将创建 local_cell 的竞争机制，enhance large freeback weight and restrain small freeback weight，更好的体现模型的收敛性，增强了模型的泛化能力。

4.2.2 AlexNet 的计算

AlexNet 是典型的卷积网络结构，包括五层卷积层、三层池化和三层全连接。论文[20]中已经说明，减少任何一个卷积层都会使得结果变得很差，所以本毕设按照该网络框架进行实现。下面具体讲解每一层的构成。

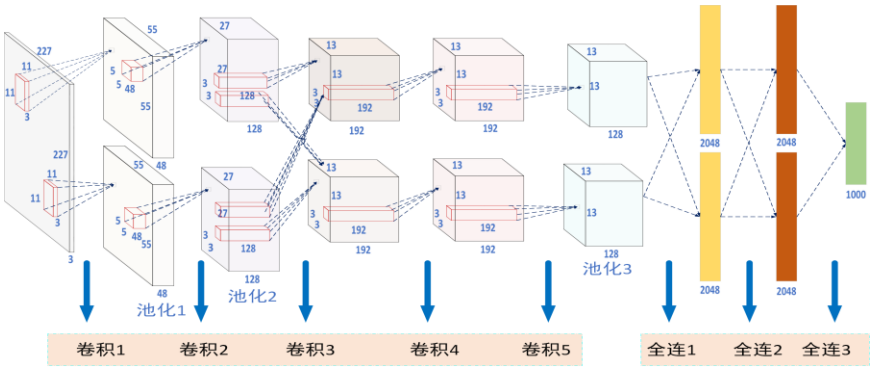


图 4. 4 AlexNet 框架结构

（1）Conv1 – relu – Max pooling – LRN，GPU=2

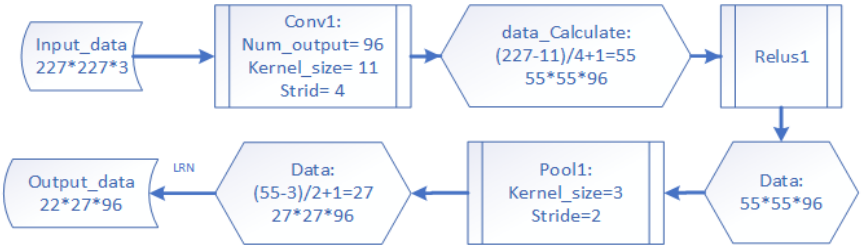


图 4. 5 Conv1 流程

（2）Conv2 – relu – Max pooling – LRN，GPU=2

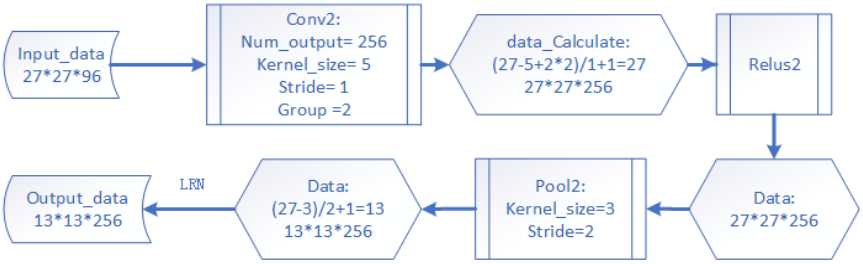


图 4. 6 Conv2 流程

（3）第三层、第四层都没有局部响应归一化和池化，只有第五层有池化。

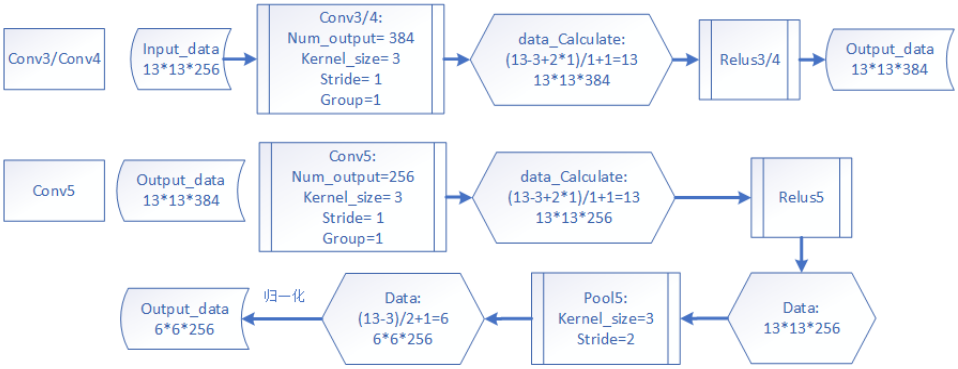


图 4. 7 Conv3/4/5 流程

（4）第六层 fc 在第五层的基础上得到 4096 维向量,再经过 dropout 层,第七层做第六层相同的操作，最后到第八层输出 softmax 为 7 的分类结果。

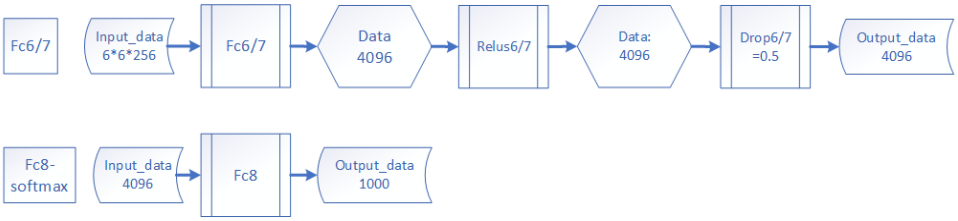


图 4. 8 Fc6/7/8 全连接

（5）本文用到的 AlexNet 框架参数如下表所示：

表 4. 1 AlexNet 框架参数列表

Input	Layer	Kernel	Stride	padding	output
n*227*227*3	C1	96*11*11*3	4	valid	n*55*55*96
n*55*55*96	Pool1	3*3	2	valid	n*27*27*96
n*27*27*96	C2	256*5*5*48	1	valid	n*27*27*256
n*27*27*256	Pool2	3*3	2	valid	n*13*13*256
n*13*13*256	C3	384*3*3*256	1	same	n*13*13*384
n*13*13*384	C4	384*3*3*192	1	same	n*13*13*384
n*13*13*384	C5	256*3*3*192	1	same	n*13*13*256
n*13*13*256	Pool3	3*3	2	valid	n*6*6*256
n*6*6*256	Fc1	4096*6*6*256	--	--	n*4096
n*4096	Fc2	4096*4096	--	--	n*4096
n*4096	Fc3	4096*7	--	--	n*7

4.3 时间金子塔结构池化

在 AlexNet 网络中，输入为单张图片，也就是每次只是输入了 25ms 的语音进行训练。每 64 帧对应一个标签，但是没有考虑空白帧和噪声帧的影响，也忽略了语音的连贯性。因此，对 DCNN 特征提取中 Fc2 全连接层的输出特征 n*4096，通过 DTPM 将其组合成固定的 7*4096。在通过 SVM 进行分类，意味着 SVM 输入为每段语音所对应的（7,4096）维特征向量，输出是七种情感的置信标签。

DTPM 是在特征中嵌入时间线索并找到最佳的集中策略。它的部分灵感来源于空间金字塔匹配（SPM）[14]，它在图像分类的特征池中嵌入了空间线索。在 SPM 中，首先将图像划分为不同尺度的区域，然后对每个区域进行特征合并。因此最

终的特征是每个尺度上的汇集特征的连接。类似地，我们还将段级特征 X 沿着时间轴以不同的比例划分为不重叠的子块。最后的连接特征因此集成了不同尺度上的时间线索，最后通过 SVM 分类器。

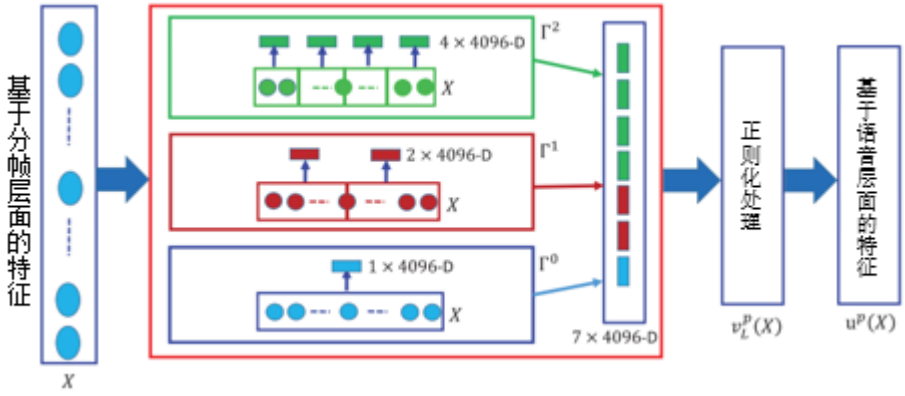


图 4. 9 DTPM 时间池化

DCNN 的全链接层输出为 Fc2 的 n*4096（n=X）维特征向量，DTPM 在多个级别上划分段级别特征 X，将特征向量 X 沿着时间轴以不同的等级被等分为 $2^{\ell}, \ell = (0,1,\cdots,L)$ 个连续的不重叠的子块：

$$X = (X_1, X_2, \cdots, X_m)$$
$$m = 2^{\ell}, \ell = (0,1,\cdots,L) \quad (4.1)$$

其中，当 $(m = 2^{\ell}) < 4$ 时，我们认为此处的段级别特征无用，不进行连接。当 $(m = 2^{\ell}) > 4, and, (m/2) \neq 0$ 时，通过最大池化后分别得到维度为 1*4096、2*4096、4*4096 的特征，采用递归的方式划分 $\Gamma^0 \xrightarrow{1/2} \Gamma^1 \xrightarrow{1/2} \Gamma^2$，以保证最后连接成 7*4096-D 维的特征向量。

4.4 SVM 支持向量机

（1）SVM 分类

在语音情感识别领域，高斯混合模型（GMM）和隐马尔科夫模型（HMM）都取得了很大的成功，但 GMM 和 HMM 更适合处理语音连续信号的问题，而 SVM 更适合实现分类效果，并且 GMM 和 HMM 都要被最大似然准则约束，Output_lable 是 same_simple 的相似度，classification 效果较差；SVM_classfication 的 Output_lable 是 different_classfication 的差异性，SVM 更加适配于不同情感特征的语音识别。

linear unseparable ：（1）将 7*4096-D feature mapping higher space，加入 Lagrange 因子α，融合 constraint condition 和 aim funcation，进行计算；（2）通过引入核函数$K(x,z)=\langle\phi(x)\bullet\phi(z)\rangle$，将 n 维的内积在低维空间中进行计算。

不断调节超平面$f(x)=w^Tx+b$中的权重 w 和偏置 b，使得分类数据到超平面的距离和最小，完成模型的训练。

（2）K-CV 交叉验证

在模型训练测试中，由于数据集的单一性，可能使得模型出现过拟合和欠拟合的现象。因此，为了减弱上述情况，论文采用交叉验证[14]的方式进行模型泛化能力的验证。

论文根据 EMO-DB 数据库中的数据按表演者的编号将原始语音数据分为 10 组，将每一组数据轮流作为模型的 Input_data，如此迭代 10 次，用测试集分类的平均置信度作为交叉分类最终的性能指标。

通过交叉验证模型，提高模型的泛化能力，得到较优的模型参数.pb 文件，和 7 中 emotion 预测的平均置信度。

4.5 代码框架

表 4. 2 python 代码结构图

File	*Funcation*
data	存放柏林语音数据库和 MFCC 图片
snaps	存放 TensorFlow 日志文件
finetune_alexnet	存放模型训练过程中保存的参数文件
alexnet.py	定义 AlexNet 网络框架
datagenerator.py	导入柏林语音数据库和 MFCC 图片，存储为 List
finetune.py	训练 AlexNet 网络模型，不断更新网络参数
test_alex.py	对 n*4096 维的特征向量，再次提取特征，将其固定成 7*4096 维特征
svm_clf.py	SVM 对每段语音的 7*4096 维特征进行分类，保存 SVM 的相关参数
classifier.py	测试用户单个语音的情感信息

第 5 章 基于 Android 端的模型部署

本章主要介绍如何将模型集成到 Android 客户端，实现简单的情感识别。

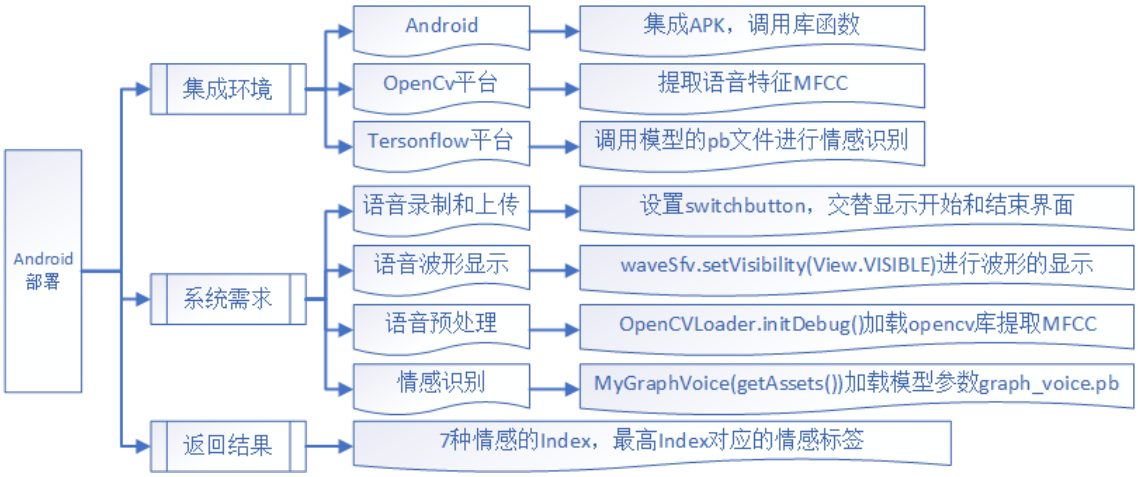


图 5. 1 情感模型的 Android 端部署实现

5.1 功能模块需求

语音情感识别系统在 Android 上部署需要包括三个模块：

（1）录制及上传: 用户实时录制语音并显示语音波形，同时用户也可以选择不进行现场录音，选择语音文件上传。

（2）语音特征提取：用户录制或上传语音后，系统能够对语音进行预处理，并提取语音特征。

（3）语音情感识别：系统识别语音情感，并返回七种语音情感类别的置信度和情感识别结果。

5.2 Android 环境配置

（1）开发平台：Android Studio

Android 环境下的开发方便与手机端的连接，集成为 APK 安装在手机上，通过手机终端更好的实时测试；并且系统部署到手机终端上，将有利于推广和使用。

（2）集成环境：Tensorflow，OpenCV

Tensorflow 用于训练模型，调用 pb 文件识别语音情感。

OpenCv 用于图像处理，因为我们将音频文件提取语音特征 MFCC，按照 RGB 三原色格式进行转换的。

运行环境：支持 android5.0 以上的系统

5.3 UI 功能实现

5.3.1 录音及语音上传

a. 录音：使用 AudioRecoed wav speech，使用 one channel and sampling rate 16Khz，encode_framing _16bit。

录音功能中主要使用了三个方法：

（1）设置了 switchbutton，用于开始和结束录音的显示按键，点击时一个显示，一个隐藏，按键进行切换。

（2）采用方法 WaveCanvas.isRecording 绘制录音过程中语音的波形，并且通过方法 waveSfv.setVisibility(View.VISIBLE)进行波形的显示。

（3）采用方法 SimpleDateFormat("HH_mm_ss")，时分秒的方式进行命名。

（4）存储格式包含了 pcm 和 wav 格式的文件。

下图是录音界面的展示：

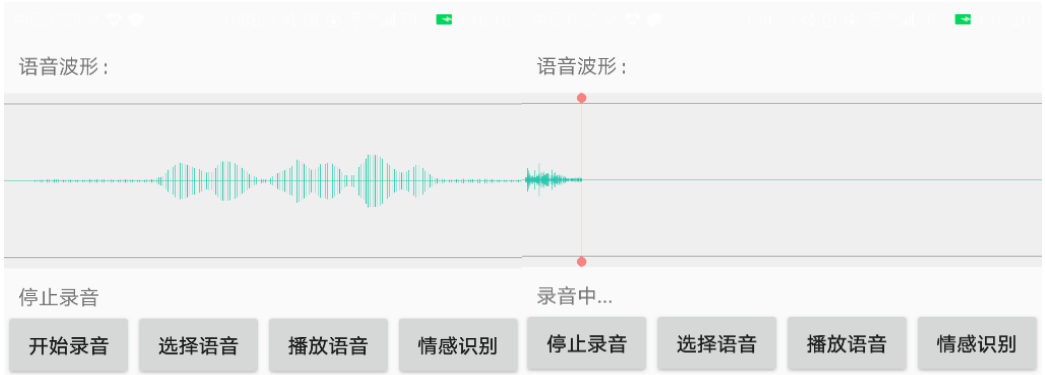


图 5. 2 Record speech UI

b. 语音上传：调用 Android 端 system 的 file manage，选择相应的 wav 文件，并且返回文件路径。

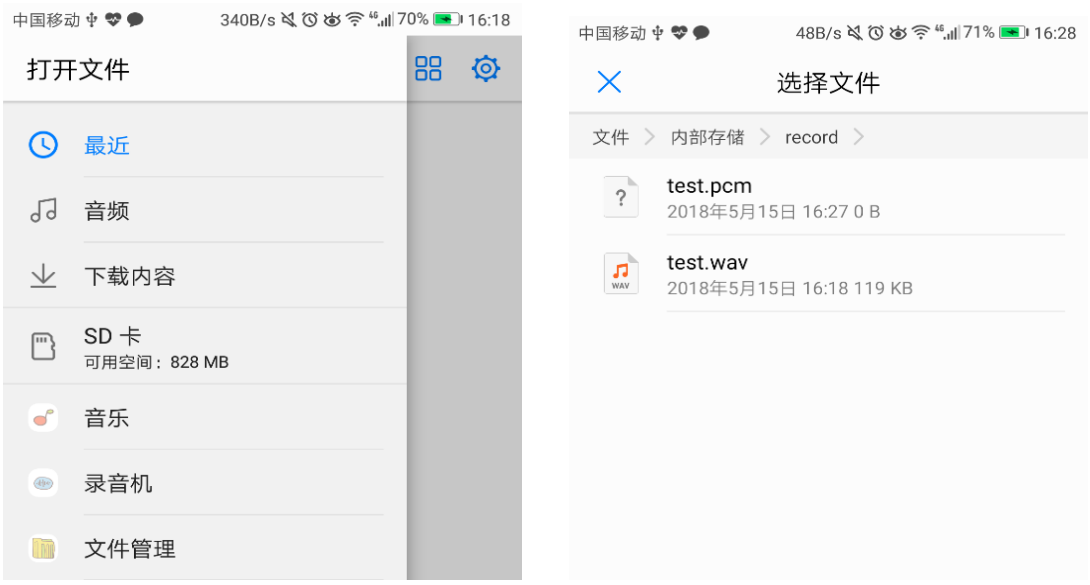


图 5. 3 File manage UI

5.3.2 语音特征提取

由于我们搭建语音情感识别系统时，采用的是 matlab 进行语音特征的提取，没有相关的库，因此没有集成到 Android，通过 Android 客户端进行的情感识别结果可能会与真实预测结果有一定出入。

库文件中 opencvlibrary320 存储的是 opencv 中所使用的库，论文将使用方法 OpenCVLoader.initDebug(); 进行 opencv 库的加载。此处的语音特征是通过加载 openCv 库，读取 wav 音频文件，同样与 MFCC 提取过程相同，对语音进行预处理和傅里叶变换后得到语谱图，提取 MFCC 特征。提取的 MFCC 特征图保存在手机的文件目录/storage/emulated/0/voice/voice_emotion_img.jpg.

图 5. 4 MFCC 特征图

5.3.3 情感识别

方法 MyGraphVoice(getAssets())加载 TensorFlow 平台，并对其进行初始化。此时便可以通过 tersonflow 平台导入模型训练好的模型参数文件 graph_voice.pb，通过方法 TensorFlowInferenceInterface(assetManager,*MODEL_FILE*)加载 pb 参数。

加载成功之后，将 opencv 处理得得到的 MFCC 特征图作为模型的输入，进行语音情感的识别，最后输出 7 中 emotion 的 index，取最大的置信度标签作为此次情感识别的结果。

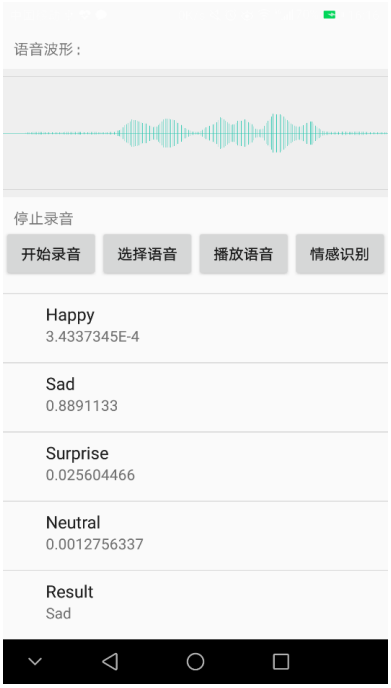


图 5. 5 情感识别结果

5.4 本章小结

本章主要讲述了语音情感识别系统在 Android 上的部署，首先简单介绍了识别系统的功能需求和开发环境，然后对部分功能的实现作了详细介绍，主要包括实时录音和时域波形显示，从本地上传语音，语音特征提取和情感识别等功能，最后进行实时语音的预测处理，得到对应情感的置信度。

第 6 章 模型测试分析

本章主要讲述的是模型的训练过程，分别对 AlexNet 框架和 SVM 框架进行训练和测试，并通过 Tersonboard 可视化训练过程；然后提出交叉验证的方法对模型进行泛化能力的验证；最后进行实时语音录制测试。

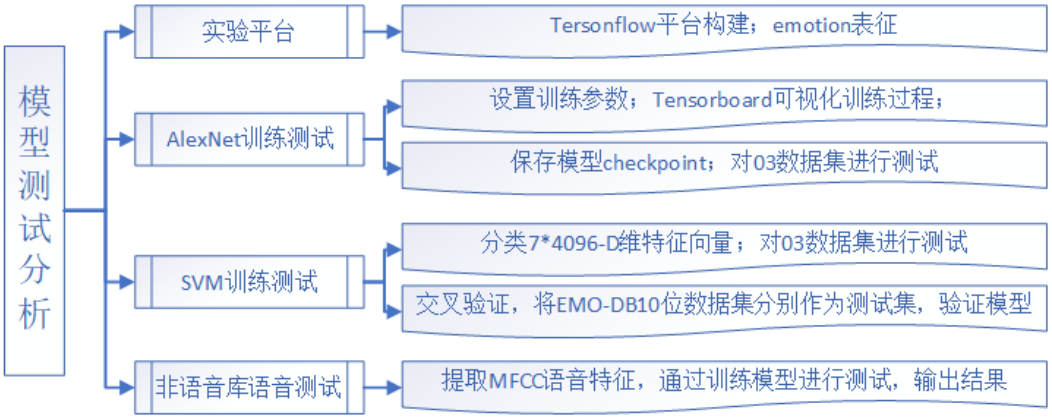


图 6. 1 模型测试框架

6.1 测试环境

本文的语音情感识别模型基于 AlexNet 框架，DTPM 和 SVM 分类算法的，训练过程是 GPU 上进行，使用 tensorflow 平台进行 Python 实现，通过 Pycharm 的 IDE 界面进行代码管理，最后通过 Tensorboard 进行结果的可视化。

实验基于 datasets 中的柏林库，trainning and test 的情感标签对应关系如下：

表 6. 1 Emotion kinds

Letter	Emotion(English)	Letter	Emotion(German)	Number
A	anger	W	Arger(Wut)	0
B	boredom	L	Langeweile	1
D	disgust	E	Ekel	2
F	anxiety/fear	A	Angst	3
H	happiness	F	Freude	4
S	sadness	T	Trauer	5
N	neutral	N	neutral	6

6.2 AlexNet 测试

6.2.1 AlexNet 训练过程

我们将 03 号表演者录制的语音作为测试集，将剩余 9 人录制的语音作为测试集，通过前向和后向反馈，对 AlexNet 框架的参数 bvlc_alexnet.npy 进行微调。微调的层数由我们指定，层数主要影响的是微调时间；微调的目标便是找到最优的 weight 和 bias，从而得到更好的训练模型。下表给出训练过程的参数：

表 6. 2 训练参数

Parameter	Value	Parameter	Value
Traning sets	08~16	Test sets	03
Learning_rate	0.001	Num_epochs	100
Batch_size	64	Dropout_rate	0.5
Num_class	7 emotion	Fine_training_layer	All_layer

通过上面的参数进行模型的训练，下图（a）显示出，AlexNet 框架模型每一层的 Input_data，图（b）显示的是每一次迭代的 Accuracy，图中是第 100 次跌带时的 Accuracy 的输出。

（a）AlexNet 框架模型

```
2018-05-21 11:09:32.977072 Epoch number: 100
2018-05-21 11:09:45.744106 Start validation
acc_val [0.45703125, 0.50390625, 0.51171875, 0.55078125, 0.5703125, 0.5859375, 0.4
9609375, 0.6015625, 0.57421875, 0.55859375, 0.56640625, 0.59765625, 0.55859375, 0.
578125, 0.5625, 0.62109375, 0.57421875, 0.6171875, 0.62109375, 0.60546875, 0.57421
875, 0.5625, 0.57421875, 0.58984375, 0.62890625, 0.58203125, 0.57421875, 0.65625,
0.58203125, 0.5703125, 0.62890625, 0.58203125, 0.625, 0.6171875, 0.57421875, 0.609
375, 0.59765625, 0.55078125, 0.58984375, 0.59375, 0.58984375, 0.609375, 0.5859375,
 0.56640625, 0.5859375, 0.59765625, 0.5859375, 0.59375, 0.6015625, 0.58984375, 0.6
09375, 0.5859375, 0.59765625, 0.58984375, 0.6171875, 0.62109375, 0.59375, 0.585937
5, 0.5703125, 0.59765625, 0.6015625, 0.59375, 0.57421875, 0.59375, 0.59375, 0.625,
 0.58203125, 0.6171875, 0.59765625, 0.6328125, 0.59765625, 0.609375, 0.6171875, 0.
6171875, 0.61328125, 0.578125, 0.609375, 0.6171875, 0.58203125, 0.6171875, 0.60937
5, 0.6171875, 0.61328125, 0.58203125, 0.6328125, 0.63671875, 0.61328125, 0.6015625
, 0.6328125, 0.6171875, 0.61328125, 0.62890625, 0.63671875, 0.609375, 0.60546875,
0.62890625, 0.609375, 0.609375, 0.63671875, 0.59765625]
2018-05-21 11:09:46.215538 Validation Accuracy = 0.5977
2018-05-21 11:09:46.215604 Saving checkpoint of model...
```

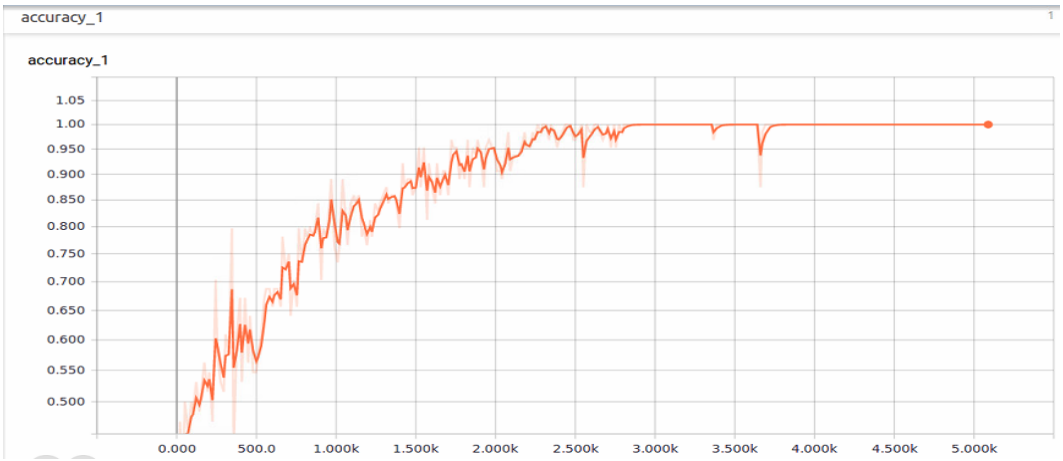
（b）epoch_number =100 时的 Accuracy

图 6. 2 AlexNet 训练过程

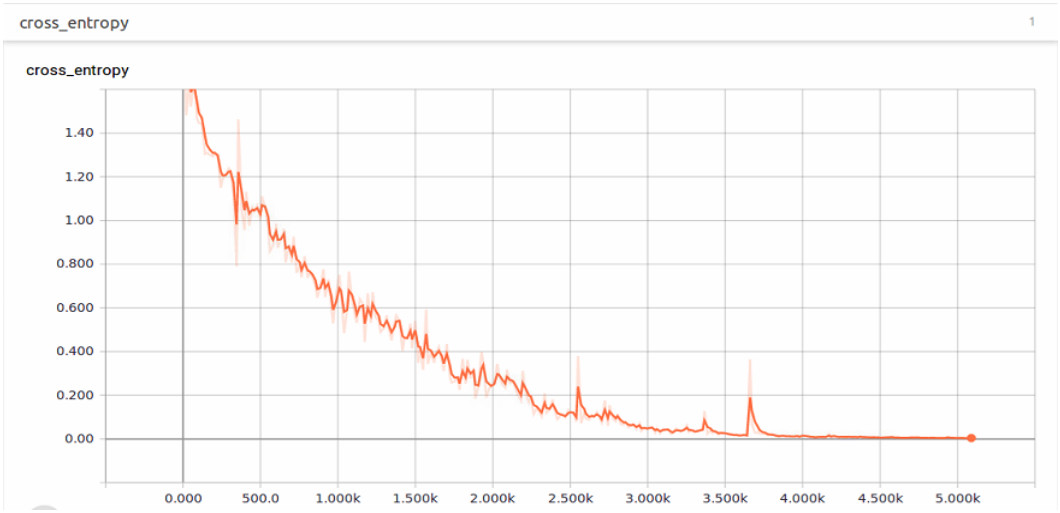
6.2.2 Tensorboard 可视化

为了直观的显示深度神经网络中的运算，我们将 tensorflow 保存 summary data 日志文件作为 TensorBoard 的输入，显示 AlexNet 训练过程的 graph，和参数 weight、bias 调节的可视化。

论文中的 tensorboard 输入是数据迭代 100 次后的结果，所有的可视化参数都是基于此文件，我们此处只显示训练过程中的 accuracy 和损失函数 cross_entropy，其余训练层的图文件，见展示过程。



（a）Training_accuracy



（b）Cross_rntropy

图 6. 3 AlexNet 训练过程可视化

下图是 03 作为测试集，将其带入模型后的 Val_accuracy，由图中的结果可以看出，直接通过 AlexNet 框架训练得到的模型，准确率只有 65%左右，不是我们所需要的最优模型。

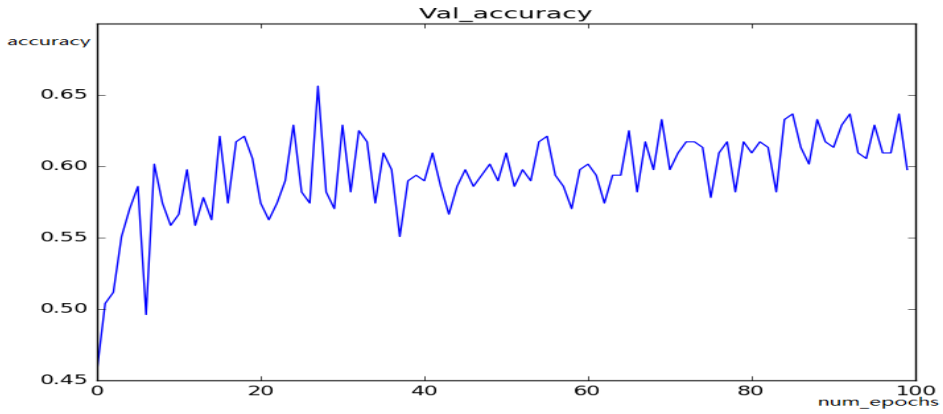


图 6. 4 AlexNet 测试结果准确

6.3 SVM 训练

通过上面的结果，如果只使用 AlexNet 训练得到模型进行测试的话，其语音识别的效果不是最佳的。因此，我们在模型的 AlexNet 模型的训练过程中进行改进，将 Fc7 全连接层的输出的 n*4096-D 的维特征向量，进行 DTPM 时间金字塔池化，通过嵌入空间线索，将每一个段级特征依次划分为 4/2/1，三种特征形式，并将其

连接成一个新的 7*4096-D 的语音特征向量，这样能集成不同尺度上的时间线索。将其作为 SVM 分类器的数据输入，进行情感标签的分类。

6.3.1 SVM 训练结果

在 SVM 分类器中，使用 sklearn.svm 库进行非线性数据的分类，下面是通过 SVM 训练后，对 03 数据集进行测试的准确率。

```
quronghui@ubuntu-X299-UD4-Pro:~/Alex_TF/finetune_alexnet$ python svm_clf.py
Test sets length:   49
************************************************************
     training.............

(train) Label  :  [6 6 0 3 2 4 6 0 0 3 2 4 5 6 3 2 4 5 6 3]
(train)Predict :  [6, 6, 0, 3, 2, 4, 6, 0, 0, 3, 2, 4, 5, 6, 3, 2, 4, 5, 6, 3]
(train)Accuracy:  0.98353909465

     testing.............
(test) Label  :  [6 2 4 6 6 2 4 5 6 2 3 2 4 3 6 6 2 2 4 3 6 3 6 2 4 0 2 3 2 4 5 6 0 3 4 5 6
 4 5 6 3 4 5 6 6 0 3 4 4]
(test)Predict :  [6 2 4 6 6 2 4 5 6 0 3 2 4 5 6 6 0 2 4 5 6 3 2 2 4 6 6 3 2 4 5 6 0 3 4 5 6
 4 5 6 3 4 5 6 6 0 1 4 4]
(test)Accuracy:  0.836734693878
************************************************************
```

图 6. 5 SVM 训练和分类结果

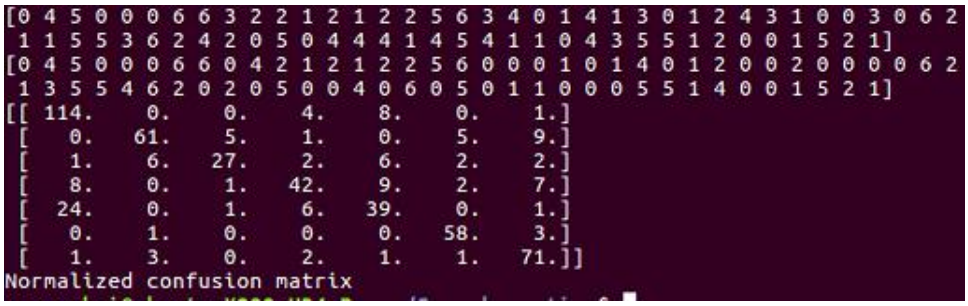
从上图可以看出，当语音特征，进过 DTPM 和 SVM 分类后，预测结果提高到了 83.6%左右。但是，只一个测试集进行模型泛化能力的检测，当测试的效果不理想的时候，可能出现不断选择训练集/测试集，使得泛化能力满意为止的情况，显然这样得到的模型并不具有价值性。

6.3.2 交叉验证

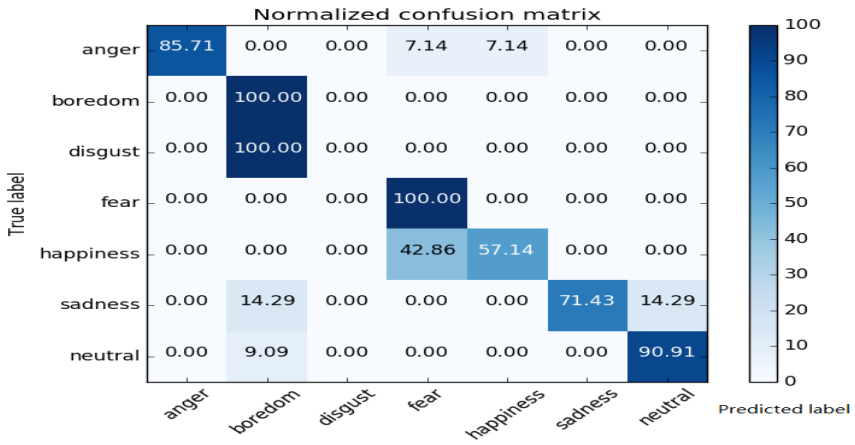
下图分别是 K=1 和 K=10 的 confusion matrix，和通过归一化处理得到的模型预测各类情感的准确率。

```
quronghui@ubuntu-X299-UD4-Pro:~/Speech_emotion$ python plot_confusion_matrix.py
[6 6 1 6 0 0 6 3 3 4 0 0 4 1 4 2 0 0 6 6 1 0 6 0 4 1 5 0 5 6 4 4 6 3 0 5 6
 0 3 6 5 5 0 0 5 5 4 1 0]
[6 6 1 6 0 0 6 3 3 3 3 0 3 1 4 1 0 0 1 6 1 0 6 0 3 1 1 0 5 6 4 4 6 3 0 5 6
 0 3 6 5 5 0 4 5 6 4 1 0]
[[ 12.   0.   0.   1.   1.   0.   0.]
 [  0.   5.   0.   0.   0.   0.   0.]
 [  0.   1.   0.   0.   0.   0.   0.]
 [  0.   0.   0.   4.   0.   0.   0.]
 [  0.   0.   0.   3.   4.   0.   0.]
 [  0.   1.   0.   0.   0.   5.   1.]
 [  0.   1.   0.   0.   0.   0.  10.]]
Normalized confusion matrix
```

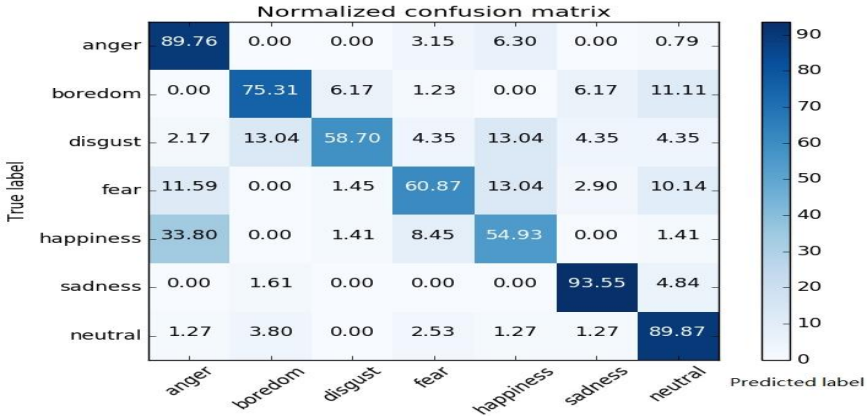
（a）K=1 confusion matrix



（b）K =10 confusion matrix



（c）K=1 Accuracy_Value



（b）K =10 Accuracy_Value

图 6. 6 交叉验证

通过交叉验证得到的实验结果：体现出模型对 7 种情感的预测的置信度，经过十次数据集的交叉训练测试，模型对测试集的平均置信度达到 74.7%，相教于 03 测试集的平均置信度 83.6%有所降低，但是为了提高模型的泛化能力，我们只能牺牲部分置信度，取最后一次。

另外一方面可以看出，上图中发现大部分的语音情感预测为：anger、sadness 和 Neutral。原因可能是提取 MFCC 基础声学特征时，语音的音高和频率刚好和这几种情感相近。所以，语音情感特征的提取需要进一步的完善。

6.4 自然录制语音测试

通过对 AlexNet 框架的微调，DTPM 的时间池化和 SVM 的情感分类，以及最后的交叉验证，我们得到了十次交叉训练得到的模型参数文件.pb。但是上面测试的过程是在 MO-DB 语音库测试得出的，论文将模型泛化到非语音库的语音测试。

通过录音工具，录制一段自然状态下的 happy.wav 语音文件，提取 MFCC 特征图片，通过模型进行测试。结果如下图，取置信度最高的标签作为预测结果。

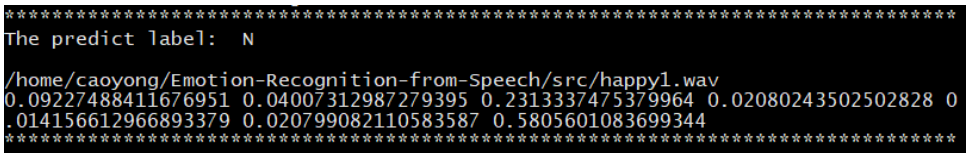


图 6. 7 自然条件下语音测试

实验测试自然条件下的语音测试，最高置信度只能达到 58%左右，相对于测试集的平均置信度，论文对于单个语音的测试取的是最高置信度，因此对自然录制语音的测试鲁棒性不是太好。

根据文献[18]，对自发 BAUM-1s 视听数据，八种情感的测试的平均置信度达到 44.61%，相较于 EMO-DB 数据集[21]的 87.31%。可以得出无论是本文的模型还是文献中的模型，对于自然环境下录制的语音，情感的识别效果相对较差。因此，如何提高自然环境下录制的语音识别的准确度，将是下一步的工作。

第 7 章 总结与展望

7.1 毕设总结

当前人工智能发展非常迅速，各类算法也在不断创新和取得突破。但是终究是基于一些特定的规则和任务，依靠大数据来进行预测和模拟。比如人脸识别，基于大数据和机器网络的学习，可以达到 90%以上的识别准确度。但是这类技术总归是机器被动地完成相应的任务，随着技术的发展，人类肯定不会只满足于这样的被动交互，机器主动和人进行交互，读懂人的情感是关键的一步。

本文主要研究了语音信号的梅尔频率倒谱系数特征图 MFCC，特征图是将语音信号转化为对应的图片，而对图片的分类或者说特征提取，最常用的是卷积神经网络。本文主要的贡献有：

（1）训练卷积神经网络进行语音情感分类，单独使用 AlexNet 效果不是很好，准确率只能达到 60%左右。因此，考虑基于 AlexNet 框架，融合 DTPM 池化和 SVM 分类器，进行情感的识别分类，模型的测试准确率达到 83.6%。

（2）尝试将算法部署到了安卓客户端上，通过 opencv 提取 MFCC 语音特征，TensorFlow 调用模型参数.pb 文件，并进行了简单的测试。当以后数据集和网络框架成熟后，可以将应用推广到更多客户端，实现人机主动交互。

通过本次毕设，初次涉猎机器学习相关的知识。从最初对数据预处理 MFCC 的提取，到学习 Python 语言和机器学习的基本原理。这个过程是艰难而充实的，同时学到了很多有趣的知识，这些算法不在只是停留在学术界，而是真正可以解决我们日常生活中的实际问题。

7.2 未来展望

本文通过对于语音信号的处理和机器学习网络框架的搭建，识别准确率均达到 80%左右。但是在实际测试的过程中，还是发现颇多不足和待解决的问题。未来如果还有机会研究本课题，可以就以下几个方面进行改进：

收集大量语音数据进行模型训练；针对不同语言的数据集训练特定模型；提高 Android 端人机交互的体验等。

参考文献

[1] 韩文静,李海峰,阮华斌,马琳.语音情感识别研究进展综述[J].软件学报,2014,25(01):37-50.

[2] Moriyama T,Ozawa S. Emotion recognition and synthesis system on speech [C]. Multimedia Computing and Systems,1999.IEEE International Conference on IEEE,1999,1:840-844.

[3] 储云云. 语音特征提取及其情感识别的研究[D]. 浙江理工大学, 2014.

[4] Jianhua Tao,Yongguo Kang.Feature Importance Analysis for Emotional Speech Classification[J].ACII 2005,LNCS 3784,2005:449-457.

[5] 赵力,黄程韦,邹采荣,余华,王开.基于情感对特征优化的语音情感分类方法中国专利, CN101894550A[P].2010-11-24.

[6] 何川.语音情感的特征提取与识别[D].华中科技大学,2008.

[7] 尤鸣宇. 语音情感识别的关键技术研究[D].浙江大学,2007.

[8] Ekman,P.AnAugment for Basic Emotions[J]. Cognition and Emotion，992，6：169-200.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss,“A database of German emotional speech.” in Interspeech, vol. 5, Lisbon,Portugal, 2005, pp. 1517–1520.

[10] Y. Wang and L. Guan, “Recognizing human emotional state from audiovisual signals*,” IEEE Transactions on Multimedia, vol. 10, no. 5, pp. 936–946, 2008.

[11] 高耀东, 侯凌燕, 杨大利. 基于多标签学习的卷积神经网络的图像标注方法[J]. 计算机应用, 2017, 37(1):228-232.

[12] 周飞燕，金林鹏，董军. 卷积神经网络研究综述[J]. 计算机学报，2017, 40(6):1229-1251.

[13] 刘万军, 梁雪剑, 曲海成. 不同池化模型的卷积神经网络学习性能研究[J]. 中国图象图形学报, 2016, 21(9):1178-1190.

[14] 丁世飞, 齐丙娟, 谭红艳. 支持向量机理论与算法研究综述[J]. 电子科技大学学报, 2011, 40(1):2-10.

[15] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The eNTERFACE’05 audiovisual emotion database,” in 22nd International Conference on Data Engineering Workshops, Atlanta, GA, USA, 2006.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss,“A database of German emotional speech.” in Interspeech, vol. 5, Lisbon,Portugal, 2005, pp. 1517–1520.

[17] Zhang S, Zhang S, Huang T, et al. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching[J]. IEEE Transactions on Multimedia, 2017.

[18] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, “BAUM-1: a spontaneous audio-visual face database of affective and mental states,” IEEE Transaction on Affective Computing, 2016.

[19] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech communication, vol. 53, no. 5, pp. 768–785, 2011.

[20] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, “BAUM-1: a spontaneous audio-visual face database of affective and mental states,” IEEE Transaction on Affective Computing, 2016.

[21] Y. Sun, G. Wen, and J. Wang, “Weighted spectral features based on local Hu moments for speech emotion recognition,” Biomedical Signal Processing and Control, vol. 18, 80–90, 2015

[22] Shiqing Zhang ,Shilliang Zhang. “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching” IEEE Early Access Articles. 2017.

[23] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, 1533–1545, 2014.

致谢

2014 年来到重庆邮电大学，我已经度过南山的近 4 个春夏秋冬，时间已逝，终究要想人生旅途中的一站说再见的时候了。4 年的学习和生活，因为有了可爱的同学们和敬爱的老师们，我的大学生活才会如此充实。此刻，对您们表示由衷的感谢。

首先，感谢我的指导老师谢昊飞老师。谢老师是一位对学术细致严谨、追求完美的老师。从大二的选修课中我和好哥们邱同学，有幸结识谢老师，并在大三的时候进入谢老师实验室，跟随谢老师进行学习。谢老师在生活和学习上给予了很多指导和帮助。在论文的撰写和修改中，谢老师一丝不苟的审阅着我的文章，知道我的文章达到您的要求。在此，学生想对您说一句：老师，您辛苦了。

其次，感谢一直陪伴我们度过四年的辅导员旭姐，是您对我们生活的指导，让我们能尽快适应大学生活，将更多的时间犹如到学习中。感谢您与我们携手相伴的四年青春。

然后，14 级的同班同学们，和寝室的好哥们，四年时光里我们一起卧谈，一起做过的那些事，我们终将铭记。感谢您们在寝室文化中的理解和支持，这终将是我最难忘的友谊。

当然，我最要感谢的是背后默默支持我的爸妈。你们用家的温暖撑起我远航的船帆，让我能在学校里快乐成长。我会更加努力，给你们一个美好的生活。

当然，我最要感谢的是一直在背后默默支持和关心我的父母。他们不善言表，但始终做我最坚强的后盾，给我丰厚的生活保障，让我衣食无忧，使我能在青海大学心无旁骛地学习和成长。在未来，我会更加努力的学习，给他们一个美好的生活。

最后，衷心感谢答辩老师，感谢您们为我们组织答辩小组，感谢您们给出的指导意见。祝我们拥有更好的么明天！

附录 A 代码链接

（a）主要代码链接

本文代码实现主要分为三部分：matlab 提取 MFCC；python 实现 AlexNet 框架训练；Android 实现 AlexNet 模型的部署。由于代码相对太多，下文主要介绍代码名和相应的实现内容，最后附上对应的代码链接，方便下载查看。

附录 A. 1 matlab 提取 MFCC

File	Funcation
Preaccentuation.m	Wav 语音的预处理和 plot 波形显示
Hanmig.m	Wav 语音的分帧加窗处理，同样 plot 波形显示
melspectrogramcomputing.m	傅里叶变化和倒谱分析
Mel2img.m	MFCC 提取的全过程，主函数

附录 A. 2 AlexNet 框架训练和参数微调

File	Funcation
Finitune.py	AlexNet 框架训练和微调 weight、bias
Test_alex.py	DTPM 池化，国定为 7*4096 维特征向量
Svm_clf.py	对特征向量进行情感标签分类
Classifier.py	测试自然录制语音置信度
Plot_confusion_martix.py	K-cross 交叉验证，提高模型的泛化能力

附录 A. 3 Android 端模型部署

File	Funcation
MainActivity.java	Android 部署的主函数
OpenCVLoader.java	Upload opencv 库，对语音进行预处理，提取 MFCC
WaveCanvas.java	进行录音波形的绘制
MyGraphVoice.Java	Uoload tensorflow 平台，调用模型.pb 文件

（附录 A.1）https://github.com/quronghui/MFCC_matlab

（附录 A.2）https://github.com/quronghui/AlexNet.git

（附录 A.3）https://github.com/quronghui/Speech_Emotion_Android_APK.git

（b）Tersonboard 可视化 AlexNet 训练过程

本文的附录主要是 AlexNet 框架迭代后，每一步对应的 weight 和 bias，此处只放置 conv1 迭代后的参数，其余的调试模型框图见附录。

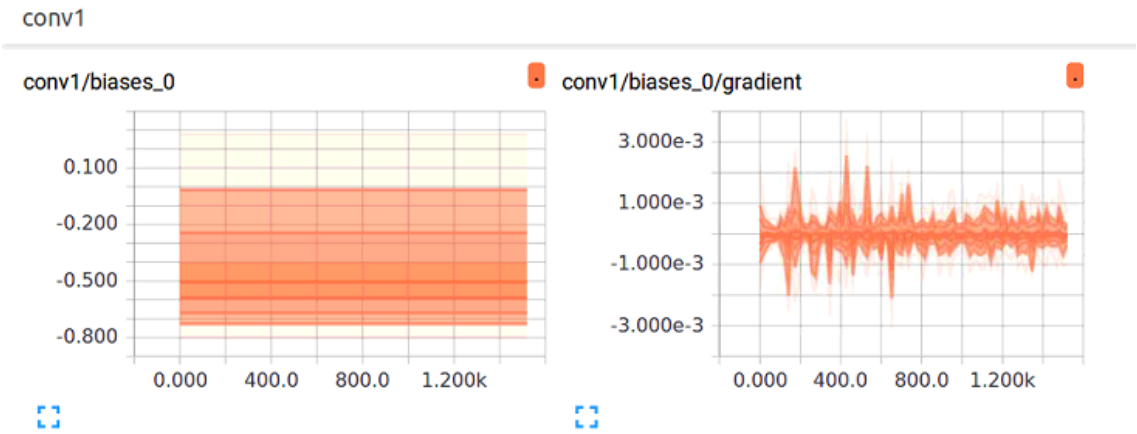


图 A. 1 Con1 bias_0 进过迭代的 bias

图 A. 2 Con1 weight_0 经过迭代的 weight

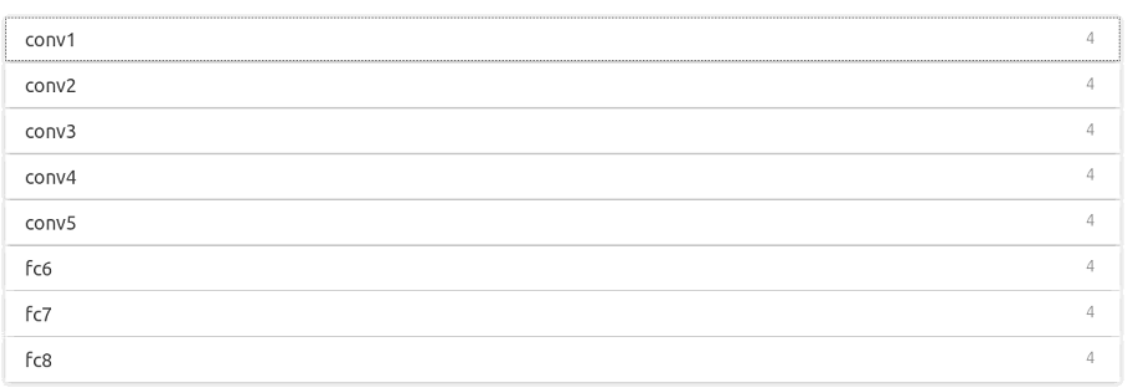


图 A. 3 AlexNet 框架 8 层结构对应的 weight 和 bias

（附录 A.4）http://192.168.3.112:6006/#distributions&_showDownloadLinks=true

附录 B 英文翻译