

# 语音情感能识别中特征抽取和选择 的研究

(申请清华大学工程硕士学位论文)

培养单位: 计算机科学与技术系  
学 科: 计算机技术  
研 究 生: 马 习  
指 导 教 师: 吴 志 勇 副 研 究 员

二〇一八年四月



# **Feature Extraction and Selection in Speech Emotion Recognition**

Thesis Submitted to  
**Tsinghua University**  
in partial fulfillment of the requirement  
for the professional degree of  
**Master of Engineering**

by  
**Ma Xi**  
**( Computer Technology )**

Thesis Supervisor : Associate Professor Wu Zhiyong

**April, 2018**



# 关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容。

本人保证遵守上述规定。

（保密的论文在解密后应遵守此规定）

作者签名：\_\_\_\_\_

导师签名：\_\_\_\_\_

日 期：\_\_\_\_\_

日 期：\_\_\_\_\_



## 摘要

在人机交互日益普遍的今天，只是理解语音中的语言学信息已经不足以满足所有需求，如何提取语音中的情感信息在许多的应用场景中也变得越来越重要。传统的语音情感能识别大致可以分为情感相关声学特征的提取和情感分类模型的构建两部分，原始语音信号通常会先被映射到情感信息相关的声学特征，然后采用各种分类模型将特征向量映射到对应的情感类别。近年来，随着深度学习方法的发展与普及，深度神经网络开始越来越多被应用到语音情感能识别，并且取得了不错的效果。此外，特征提取和情感分类两个部分也开始被整合到一起，通过深度神经网络将可以构建从原始语音信号直接到情感类别的端到端的识别系统，但是如何为不同情感更有效地抽取和选择特征并没有被现有的研究广泛关注。针对这个问题，本文将分别从传统的语音情感能识别方法和端到端的深度学习方法入手，设计对应的方法来抽取和选择特征，进而提升系统的识别率。主要的研究工作和贡献如下：

**一、提出一种基于情感对的语音情感能识别框架，为不同的情感对选择不同的声学特征，并在最后的决策融合过程中引入心理学的情感空间模型，从而提升了系统的识别率。**传统的语音情感能识别系统通常为所有的情感选取相同的声学特征来完成最后的情感分类，但实验证明不同的情感和不同的声学特征的相关性并不相同。针对这一问题，我们将分别为不同的情感对选取不同的特征集合，将原先的多分类问题转变为多个二分类问题，并在最后的决策融合过程中通过贝叶斯分类器引入情感空间的信息。在公开的情感语音数据集 IEMOCAP 上，我们方法取得了比传统的识别框架更好的准确率。

**二、构建了基于深度神经网络的端到端的语音情感能识别系统，使用语谱图代替传统的声学特征，从而提升了系统的识别率。**随着深度学习技术和工具的发展，许多的研究者开始采用深度神经网络在原始语音信号上直接构建分类或者回归模型，被称之为端到端的系统。相比于传统的声学特征，这种方法可以抽取到更符合任务目标的特征表示。语谱图是语音信号的一种无损表示，我们通过卷积神经网络从语谱图上直接抽取和情感相关的特征表示，然后通过循环神经网络来建模语音信号的时序信息，最后通过全连接网络将输出映射到不同情感的后验概率。相比于传统的语音情感能识别，端到端的系统能够取得更好的准确率。

**三、设计了一种能够处理变长语音段的神经网络结构来实现端到端的系统，消除了语音分段带来的中性语音和情感语音的混淆，从而提升了系统的识别率。**在

## 摘要

---

使用深度神经网络实现端到端的语音情感识别系统时，由于卷积神经网络和循环神经网络很难处理变长的输入，通常会把变长的语音句子切分成多个等长的语音段，然后将所有语音段都标记为对应句子的情感标签，但这样会导致部分中性语音段被标记为有情感。针对于这一问题，我们采用补齐和掩码的方式来处理神经网络中变长的输入序列，保证模型可以接受整个语音句子，从而避免了错误标注带来的效果变差。相对于切分等长语音段的方法，我们直接输入整个变长语音句子的方法可以在相同的数据集上取得更好的准确率。

**关键词：**语音情感识别；情感对；维度情感空间模型；深度神经网络；变长语音段

## Abstract

Nowadays human-machine interaction is becoming more and more popular. The machine only understands linguistic information in speech, which is not enough to satisfy all needs. How to extract emotional information in speech is becoming more and more important in many application scenarios. Traditional speech emotion recognition can be divided into two parts: acoustic feature extraction and emotion classification model. The original speech signals are usually mapped to the acoustic features, and then use the classification models to map the features into the corresponding emotional categories. In the recent years, with the development of deep learning, deep neural networks have been introduced to speech emotion recognition, and have achieved good results. Furthermore, feature extraction and emotion classification are beginning to be integrated together. Through deep neural network, the end-to-end recognition system can be built from original speech signals to emotional categories. However, how to extract and select features more effectively for different emotions has not been widely studied. In order to solve this problem, this paper will design the corresponding methods to extract and select features with traditional speech emotion recognition and end-to-end speech emotion recognition, and then improve the recognition rate of the system. The main contributions of this paper include:

**1. A emotion-pair based speech emotion recognition framework is proposed to select different features for different emotion-pairs, and the dimensional emotion space is introduced in the decision fusion, thus improving the recognition rate of the system.** In general, the traditional speech emotion recognition system selects the same acoustic features for all emotions to complete the emotion classification, but the studies have shown that the correlation between different emotions and different acoustic features is different. To solve this problem, we will select different acoustic features for different emotions, and turn the multi-classification problem into bi-classification problems. In the decision fusion stage, the information of emotional space is introduced by Bayes classifier. In the IEMOCAP emotional speech dataset, our method can achieve better accuracy than the traditional recognition framework.

**2. The deep neural network based end-to-end speech emotion recognition system is constructed, which uses the spectrogram instead of the acoustic features,**

**thus improving performance of system.** With the development of deep learning techniques and tools, many researchers have begun to use deep neural network to directly construct classification and regression models on original speech signals, the so-called end-to-end system. Compared with traditional acoustic features, this method can extract feature representations which are more related to the target. Spectrogram is a lossless representation of speech signals. We use the convolution neural networks to extract the emotion related features directly from the spectrogram, and then model the time sequence information of the speech signals through recurrent neural networks, and map the output to the posterior probability of the different emotions through the full-connected neural networks. Compared to the traditional speech emotion recognition, the end-to-end system can achieve better accuracy.

**3. A neural network structure which can handle the variable-length speech segments is designed to implement the end-to-end system. This structure can relieve the confusion between neutral speech and emotional speech brought by the speech segmentation, thus improving the recognition rate of the system.** When using deep neural networks to implement the end-to-end speech emotion recognition system, because the convolution neural network and recurrent neural network are difficult to deal with the variable-length input, the speech sentences are usually cut into multiple equal-length speech segments, and all the speech segments are labeled as the emotional labels of the corresponding sentence. However, this causes some neutral speech segments to be marked with other emotions. In order to solve this problem, we use the method of padding and masking to deal with the variable-length input sequence in the neural networks, and ensure that the model can accept the whole speech sentence. This will avoid the performance degeneration through introducing the wrong labels. Compared with the method of segmenting equal-length speech segments, we can get better accuracy on the same dataset by directly inputting the whole variable-length speech sentence.

**Key words:** Speech Emotion Recognition; Emotion-Pair; Dimensional Emotion Space;  
Deep Neural Network; Variable-Length Speech Segments

## 目 录

<b>第1章 绪论 .....</b>	<b>1</b>
1.1 研究背景和意义.....	1
1.2 研究现状.....	2
1.2.1 传统的语音情感识别.....	2
1.2.2 端到端的语音情感识别 .....	3
1.3 本文的主要研究内容和贡献.....	5
1.3.1 研究内容和各章介绍 .....	5
1.3.2 本文主要贡献 .....	6
<b>第2章 语音情感识别的相关工作 .....</b>	<b>8</b>
2.1 本章引论.....	8
2.2 情感的定义 .....	8
2.2.1 离散的情感类别标签.....	8
2.2.2 连续的情感维度空间.....	8
2.3 情感语音数据库.....	9
2.3.1 设计准则.....	10
2.3.2 常用的情感语音数据库 .....	12
2.4 声学特征的抽取.....	13
2.4.1 人工选择情感相关的声学特征.....	13
2.4.2 选择算法筛选情感相关的声学特征 .....	14
2.4.3 深度神经网络抽取情感相关的声学特征.....	15
2.5 情感分类模型的构建 .....	16
2.5.1 基于传统机器学习的情感分类模型 .....	16
2.5.2 基于深度学习的情感分类模型 .....	18
<b>第3章 基于情感对的语音情感识别框架 .....</b>	<b>20</b>
3.1 本章引论 .....	20
3.2 情感对 .....	21
3.2.1 情感对的定义 .....	21
3.2.2 基于情感对的声学特征选择 .....	22
3.2.3 基于情感对的二分类模型 .....	27
3.3 决策融合 .....	28

3.3.1 基于投票的决策融合 .....	28
3.3.2 基于情感空间的贝叶斯决策融合 .....	30
3.4 实验结果及分析 .....	32
3.4.1 实验设置 .....	33
3.4.2 实验结果 .....	34
3.5 本章小结 .....	38
<b>第 4 章 基于语谱图的端到端的语音情感能识别 .....</b>	<b>40</b>
4.1 本章引论 .....	40
4.2 基于语谱图的卷积神经网络特征抽取 .....	40
4.2.1 语谱图的定义 .....	40
4.2.2 卷积神经网路 .....	42
4.2.3 基于语谱图的卷积神经网络特征抽取 .....	44
4.3 基于循环神经网络的时间序列建模 .....	44
4.3.1 循环神经网络 .....	44
4.3.2 基于循环神经网络的时间序列建模 .....	46
4.4 实验结果及分析 .....	46
4.4.1 实验设置 .....	46
4.4.2 实验结果 .....	49
4.5 本章小结 .....	50
<b>第 5 章 基于变长语音段的情感能识别 .....</b>	<b>51</b>
5.1 本章引论 .....	51
5.2 变长语音段的语谱图抽取 .....	52
5.3 变长神经网络结构 .....	52
5.3.1 变长卷积神经网络 .....	52
5.3.2 变长循环神经网络 .....	53
5.4 实验结果及分析 .....	53
5.4.1 实验设置 .....	54
5.4.2 实验结果 .....	54
5.5 本章小结 .....	56
<b>第 6 章 总结与展望 .....</b>	<b>58</b>
6.1 本文工作总结 .....	58
6.2 未来工作展望 .....	59
<b>参考文献 .....</b>	<b>60</b>

## 目 录

---

致 谢 .....	66
声 明 .....	67
个人简历、在学期间发表的学术论文与研究成果 .....	68

## 第1章 绪论

### 1.1 研究背景和意义

语音作为人与人之间交流最为自然的一种媒介，在我们的日常生活中起着至关重要的作用，这也使得语音一直被许多研究者认为是最有效的人机交互方式。在最近的二十年间，语音识别技术已经取得飞速的发展，它的目的是将人们说出的语言转换成对应的词序列。尽管语音识别技术已经被广泛的应用，但我和真正的自然人机交互仍然有很大的距离，因为机器仍然无法理解说话人的情感状态，而语音中通常会包含说话人当时的情感信息，这使得语音情感识别技术开始受到重视，它的目的就是从语音中识别出说话人的情感。

语音情感识别在许多需要自然人机交互的应用中提供帮助，例如网页影评和计算机辅助教育这些需要用户反馈的应用场景。在车载系统中，情感识别可以通过语音检测驾驶员当前的精神状态，并在合适的时候做出提醒从而保证车辆安全。语音情感识别还可以作为心理咨询师提供诊断的辅助工具。在同声传译系统中，语音情感识别也可以通过检测说话人情感状来调整翻译内容。对于语音识别系统来说，加入情感信息同样可以提高识别准确率。还有在电话客服系统中，客户由于长时间的等待或者紧迫的需求，声音会变的烦躁和愤怒。在这种情况下，情感识别系统就可以通过检测语音中愤怒的程度来为这些客户安排服务优先级。此外，在聊天机器人中，语音情感识别系统可以通过检测对方的情感状态来做出不同的应答。这种功能同样也可以运用在玩具中，当孩子不高兴的时候，玩具就可以通过一些方式安抚他们。在电子游戏中，尤其是在一些有语音交互的游戏中，情感识别可以通过模拟更加真实的交互场景来提升游戏体验。因此，语音情感识别在近几年来收到越来越多的关注。无论是在前沿模型算法的研究，还是在相关产品的落地，很多研究机构和公司都在这一领域投入大量的精力。

语音情感识别也存在许多挑战性的问题<sup>[1,2]</sup>。首先，很难确定哪些声学特征与情感信息最相关。一些在语音识别领域公认的特征，例如梅尔频率倒谱系数 (Mel-Frequency Cepstral Coefficients, MFCC)，在语音情感识别上并没有取得好的效果。其次，一段语音中有时会包含多种情感，如何找到每种情感对应的语音边界是个很有挑战性的问题。此外，由于说话人的文化背景不同，也会导致表达情感的方式不太相，例如有的人愤怒时会说话很大声，而有的人却声音很低沉。语言同样也是影响情感表达的一个主要问题，例如汉语和德语在表达愤怒时有一些不同。还有就是情感并没有公认的规范定义，只能通过人们的主观感觉来区分，这会导致

出现许多的情感类别。一些研究者倾向于一种“调色板”理论，就是只定义几种基本的情感，然后其他的情感都是由这些基本情感以不同的比例调和而成，这些基本情感被称为原型情感。情感语音的数据相对来说也很难获取，这会导致许多复杂的模型无法得到充分训练。

尽管目前还没有非常成熟的语音情感识别的产品，但许多公司已经在极力推动这一领域的落地。Facebook早在2012年就开始进行对用户的情绪检测实验，期望能够通过情感信息来优化他们的推荐系统。微软小冰也嵌入了语音情感识别模块，希望能够和人们更自然的聊天。IBM与软银合作推出了具有情绪感知能力的机器人Pepper。苹果也开始在自己的产品中加入语音情感识别的功能。国内的公司在近几年也开始这一领域的产业化，科大讯飞公司已经开始推出语音情感识别相关的技术支持，百度视频也推出基于情感识别的内容推荐系统。此外，也有一些创业公司开始关注语音情感识别，例如竹简智能。

## 1.2 研究现状

语音情感识别根据处理流程不同大致可以分为两种类型，一种是从语音信号中抽取声学特征，然后通过分类器区分情感类别的传统方法，另一种是直接将原始语音信号映射到情感类别的端到端系统，下面将分别介绍这两类方法。

### 1.2.1 传统的语音情感识别

传统的语音情感识别主要可以分为两个部分，第一部分是抽取和情感信息相关的声学特征，第二部分是使用分类器将特征向量映射到对应的情感类别，图1.1是传统语音情感识别的流程图。

特征选择是一个在所有分类问题中都存在的问题，目的是抽取与分类目标最相关的特征子集。语音情感识别作为一个分类任务，在这方面已经有很多的研究工作。首先，韵律学相关的特征已经被广泛的应用在语音情感识别<sup>[3-5]</sup>，包括基频相关的特征，能量相关的特征和时长相关的特征。其次，谱相关的特征也在情感识别中起到了重要的作用<sup>[6-11]</sup>，例如线性预测系数(Linear Prediction Coefficients, LPC)，线性预测倒谱系数(Linear Prediction Cepstral Coefficients, LPCC)以及梅尔频率倒谱系数(Mel-frequency Cepstral Coefficients, MFCC)。此外，声音质量相关的特征也被证明同情情感识别任务相关<sup>[3,12]</sup>。

从语音信号中抽取声学特征后，接下来就变成了一个基本的分类问题，有许多的分类模型已经被运用在情感识别任务中。隐马尔可夫模型(Hidden Markov Model, HMM)是被广泛使用在语音识别中的一种模型<sup>[13-18]</sup>，它的运作原理和语音产生

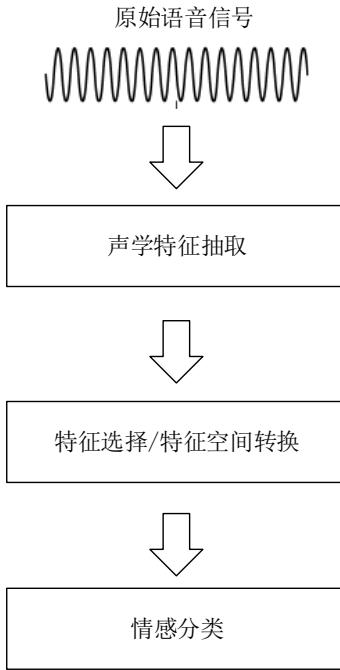


图 1.1 传统语音情感识别流程图

的机制十分相似，因此这种模型同样也在语音情感识别中被使用。高斯混合模型 (Gaussian Mixture Models, GMM) 是一种概率分布模型<sup>[19–25]</sup>，它使用多变量的高斯分布来预测当前语音句子属于不同情感类别的概率，在一些公开数据集上取得了不错的效果。支持向量机 (Support Vector Machine, SVM) 是一种被广泛使用在许多模式识别任务中的分类模型，它通过核函数 (Kernel Functions) 将低维空间无法区分的特征向量映射到更高维的空间，然后用线性分类器将其区分，这种模型在许多语音情感识别的研究中也被使用<sup>[26–28]</sup>。随着计算能力和存储容量的提升，深度学习模型在近几年变得逐渐流行，在许多的任务上都超过了传统的机器学习算法。在语音情感识别领域，深度学习的方法同样也被广泛使用<sup>[29–33]</sup>。除了上面这些单独的分类模型，还有一些工作尝试将多种模型混合使用，然后共同决策最后的情感类别，期望能够提高系统的鲁棒性<sup>[34–39]</sup>。

### 1.2.2 端到端的语音情感识别

在最近几年，深度学习的方法和工具已经被运用到语音处理领域<sup>[40–45]</sup>，包括用于特征抽取，分类和回归任务，或者两者兼而有之。一些试验结果显示在原始语音信号上使用深度神经网络提取特征可以比采用人工定义的声学特征得到更好的效果。这导致许多的研究者开始采用端到端的系统，即省略声学特征提取的过程，直接建立从原始语音信号到任务目标的映射，图1.2是端到端的语音情感识别流程

图。

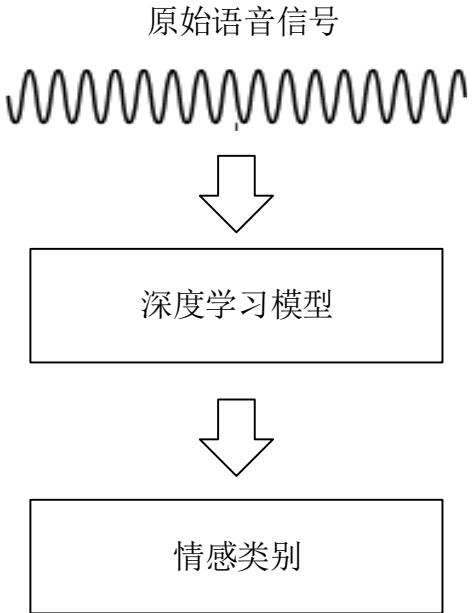


图 1.2 端到端语音情感情识别流程图

这种端到端的系统最早出现在语音识别领域，最早的工作是 Jaitly 等人<sup>[46]</sup> 通过受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 从原始语音信号上得到一种有利于语音识别的中间表示。Bhargava 等人<sup>[47]</sup> 则是通过堆叠的全连接神经网络从原始语音信号得到瓶颈特征 (Bottleneck Feature)，并且取得了和使用梅尔频率倒谱系数相近的效果。Sainath 等人<sup>[48]</sup> 将 CNN 和 LSTM-RNN 用在大词表语音识别上。Hannun 和 Amodei<sup>[49]</sup> 在线性语谱图 (Linearly-Spaced Spectrogram) 上采用深度神经网络，并搭建出了当时最好的语音识别系统。此外，梅尔语谱图 (Mel-Scale Spectrogram) 上使用深度学习方法，在说话人识别上也取得了很不错的效果<sup>[50]</sup>。

在语音情感情识别领域也开始有一些端到端的方法被提出。George 等人<sup>[51]</sup> 提出一种使用 CNN 从原始语音信号提取特征，然后通过 LSTM-RNN 捕获输入序列的时序信息并最终输出不同情感的后验概率，并且他们发现 LSTM-RNN 不同节点的输出和一些声学特征有很强的相关性。Satt 等人<sup>[52]</sup> 也采用了相似的神经网络结构，但不同的是他们从语谱图 (Spectrogram) 上抽取特征而非原始语音信号。他们认为在语谱图上可以更方便的进行去噪的操作，并且他们在公开情感情语音数据集 IEMOCAP 上取得了超过之前最好结果 (The State of Art) 的准确率。

## 1.3 本文的主要研究内容和贡献

### 1.3.1 研究内容和各章介绍

语音情感能识别效果与声学特征的选取是密切相关的，抽取什么特征以及如何抽取特征最终会影响到识别率。本文会从传统的语音情感能识别和端到端的语音情感能识别两类方法分别研究特征抽取和选择的方法，意使最终的情感能识别准确率得到提高。

第2章主要介绍语音情感能识别相关的基础知识，包括情感的定义，情感语音数据库，情感相关的声学特征以及情感分类模型这四个方面的内容。首先，我们介绍了两种不同的情感定义方式，并分析了各自的优缺点。接下来我们介绍了不同类型的情感语音数据库，包括他们的采集方式和标注方式。然后我们会对情感相关的声学特征做一个全面的介绍，包括特征类型，特征选择算法和深度神经网络抽取特征。最后我们会对不同的语音情感分类模型做出介绍，包括传统的机器学习模型和深度学习模型。

第3章介绍了基于情感对的语音情感能识别框架。通过将任意两种不同的情感组成情感对，并为不同的情感对分别选择不同的特征子集构建二分类器，从而得到更精确的二分类结果。进一步，我们发现在维度情感空间中，不同情感之间的距离并不相同，距离越近表示情感之间更相似，越远表示越不相似。依据这种信息，我们构建贝叶斯分类器来对所有情感对的输出结果做决策融合，从而得到最终的情感类别。在实验结果上我们超过了为所有情感的分类选取相同特征集的方法，同时我们的效果也优于基于决策树的分层识别框架，这种框架的设计思想和我们很相似，但我们的方法效果更好，而且当情感类别变化时更方便构建。

第4章介绍了基于深度学习的端到端的情感能识别方法。在这一章我们将不再采用传统的声学特征提取方法，而是转而使用深度神经网络将语谱图 (Spectrogram) 直接映射到对应的情感类别。我们会将语音信号首先转换为语谱图，然后采用 CNN 直接在语谱图上抽取特征。由于语音信号属于时间序列信号，因此在 CNN 抽取特征后将采用 RNN 建模输入序列的时序信息，并将最后一个时间步的输出传给后面的全连接网络，最后通过全连接网络将输入映射到每种情感的后验概率。相比于传统声学特征，采用语谱图作为输入可以取得更好的识别效果。

第5章介绍了如何设计能够处理变长语音段的深度神经网络结构。上一章将句子切分成多个更短的等长语音段，并将每个语音段都标记为所属句子的情感类别，因为大多数网络结构更容易处理定长的输入序列，但是将句子切分成更短的等长语音段无法保证每个语音段都包含有情感信息，这会导致网络在训练时对中性语音和情感语音产生混淆。因此，我们设计了一种能够处理变长语音段的神经

网络结构，通过补齐和掩码可以保证在训练和测试的过程中都不需要对语音段进行切分，这样就可以避免上面的问题。实验结果显示，变长方法在公开的情感语音数据库 IEMOCAP<sup>[53]</sup> 上超过了定长方法得到的最好准确率。

第6章对本文中关于语音情感识别中声学特征的抽取和选择的相关研究成果进行了总结，同时，还对基于情感对的识别框架和基于深度学习的端到端的框架的使用前景进行了展望。

### 1.3.2 本文主要贡献

本文的主要贡献点有以下几个方面：

**一、提出一种基于情感对的语音情感识别框架，为不同的情感对选择不同的声学特征，并在最后的决策融合过程中引入心理学的情感空间模型，从而提升了系统的识别率。**传统的语音情感识别系统通常为所有的情感选取相同的声学特征来完成最后的情感分类，但实验证明不同的情感和不同的声学特征的相关性并不相同。针对这一问题，我们将分别为不同的情感对选取不同的特征集合，将原先的多分类问题转变为多个二分类问题，并在最后的决策融合过程中通过贝叶斯分类器引入情感空间的信息。在公开的情感语音数据集 IEMOCAP 上，我们方法取得了比传统的识别框架更好的准确率。

**二、构建了基于深度神经网络的端到端的语音情感识别系统，使用语谱图代替传统的声学特征，从而提升了系统的识别率。**随着深度学习技术和工具的发展，许多的研究者开始采用深度神经网络在原始语音信号上直接构建分类或者回归模型，被称之为端到端的系统。相比于采用传统的声学特征，这种方法可以抽取到更符合任务目标的特征表示。语谱图是语音信号的一种无损表示，我们通过卷积神经网络来从语谱图上直接抽取和情感相关的特征表示，然后通过循环神经网络来建模语音信号的时序信息，最后通过全连接网络将输出映射到不同情感的后验概率。相比于采用传统的声学特征，端到端的系统能够取得更好的准确率。

**三、设计了一种能够处理变长语音段的神经网络结构来实现端到端的系统，消除了语音分段带来的中性语音和情感语音的混淆，从而提升了系统的识别率。**在使用深度神经网络实现端到端的语音情感识别系统时，由于卷积神经网络和循环神经网络很难处理变长的输入，通常会把变长的语音句子切分成多段等长的语音段，然后将所有语音段都标记为对应句子的情感标签，但这样会导致部分中性语音段被标记为有情感。针对于这一问题，我们采用补齐和掩码的方式来处理神经网络中变长的输入序列，避免了错误标注带来的效果变差。相对于切分等长语音段的方法，我们直接输入整个变长语音句子的方法可以在相同的数据集上取得更

好的准确率。

## 第2章 语音情感识别的相关工作

### 2.1 本章引论

目前，语音情感识别方面的研究工作越来越多。由于情感感知的人为主观性比较强，这导致当前的研究不只是模型算法方面的研究，还包括心理学方面的研究和人文社会学方面的研究。此外，情感语音数据库的采集和标注也是一个很大的挑战。本章将对语音情感识别的相关工作进行简单的介绍，其中包括情感的定义、情感语音数据库、声学特征的抽取以及情感分类模型的构建四个方面。

### 2.2 情感的定义

情感在心理学上的定义为：“人对客观现实的一种特殊反映形式，是人对于客观事物是否符合人的需要而产生的态度的体验”<sup>[54]</sup>，但这种定义太过宽泛，在实际的语音情感识别任务中无法运用，只有将情感通过数学量化表示后才能够被模型处理。目前对于情感的主流定义方式大致分为两种，分别是离散的情感标签定义和连续的情感维度空间定义，下面将分别对这两种定义进行介绍。

#### 2.2.1 离散的情感类别标签

在我们的日常生活中，我们在描述自己的主观感受时，通常会用一些特定的词汇，例如高兴，愤怒，悲伤等等。在情感识别中通常会将这些词汇作为情感的类别，进而将任务转化为多分类问题。关于具体应该将情感分为那些类别，不同的学者有着不同的定义和划分，下面的表格2.1列举了不同的定义方式<sup>[55,56]</sup>。

这种定义情感的方式的优点是简单、易懂，而且可以有比较明显的应用场景，这也使得当前关于情感识别的研究主要都是基于这种定义进行的，本文也主要是基于这种情感定义展开的。但缺点是对情感的描述能力有限，因为情感标签的描述太过模糊，对于一些复杂的情感无法准确的定义。例如，愤怒可以分为冷愤怒和热愤怒；又比如，高兴又可以分为不同的等级，从喜上眉梢，到眉飞色舞，再到手舞足蹈。

#### 2.2.2 连续的情感维度空间

另外一种情感的定义方式是连续的情感维度空间定义<sup>[3]</sup>，这种定义将情感映射到一个笛卡尔空间坐标系中，不同的坐标轴分别代表不同的心理学属性，每一

表2.1 不同学者对情感的定义

学者	情感类别
Arnold	愤怒, 厌恶, 无畏, 忧郁, 渴望, 绝望, 珍视, 憎恨, 希冀, 爱慕, 悲伤
Ekman, Friesen, Ellsworth	愤怒, 厌恶, 恐惧, 高兴, 悲伤, 惊讶
Fridja, Gray	希冀, 高兴, 有趣, 惊讶, 渴望, 悲伤
Izard	愤怒, 轻蔑, 厌恶, 悲伤, 恐惧, 内疚, 有趣, 高兴, 羞愧, 惊讶
James	恐惧, 悲伤, 爱慕, 愤怒
McDougall	恐惧, 厌恶, 高兴, 顺从, 柔和的情感, 渴望
Oatley, Johnson-Laird, Panksepp	愤怒, 厌恶, 焦虑, 高兴, 悲伤
Plutchik	认可, 愤怒, 希冀, 厌恶, 高兴, 恐惧, 悲伤, 惊讶
Tomkins	愤怒, 有趣, 轻蔑, 厌恶, 悲伤, 恐惧, 高兴, 羞愧, 惊讶
Watson	恐惧, 爱慕, 愤怒
Weiner, Graham	高兴, 悲伤

种情感都可以被视为坐标系中的一个点。常用的情感空间模型有二维情感空间(激活度-效价)和三维情感空间(激活度-效价-支配力),下面的图2.1和图2.2分别是两种情感维度空间的示意图。

这种情感空间模型理论上可以描述任何情感,其中激活度代表情感的强烈程度,例如愤怒就是激活度非常高的情感。效价代表情感的积极性,例如悲伤就是一种积极性很低的情感,所以它的效价很低。支配力代表情感对别人的影响程度,例如高兴就是对别人影响比较大的一种情感,所以支配力比较大。情感空间模型将原先的标签分类问题转换为的对心理学属性值的回归问题,从而能够描述更为复杂的情感。但这种模型的缺点是标记数据的成本太高,因为将主观情感量化为客观数值是一个繁重且难以保证质量的过程。

### 2.3 情感语音数据库

在大多数的监督性机器学习问题中,训练数据一直是一个很大的问题,尤其是如何得到高质量的训练数据,对于语音情感能识别更是如此。由于语音中情感信

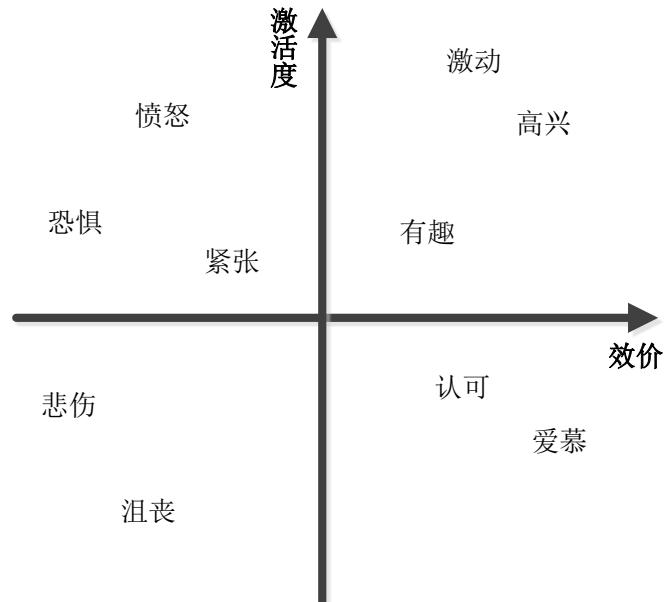


图 2.1 二维情感空间模型（激活度-效价）

息的标注主要是要依靠人来完成的，但不同的人对情感的感知程度是不同的，所以同一句话有可能被不同的人标注为不同的情感。下面的两小节将主要介绍情感语音数据库的设计准则和一些常用的情感语音数据库。

### 2.3.1 设计准则

如何判断情感语音数据库能够模拟真实的应用场景，这需要相应设计准则来指导，下面将介绍几种主要的设计准则。

自然语音还是表演式语音？通常来说，最符合实际的语音数据应该是从日常生活的对话中收集<sup>[57]</sup>，例如广播电台，电话客服系统等，这样的录音包含有最自然的情感表达。但不幸的是，由于一些法律和道德的原因，这样的数据被禁止用作研究目的。所以现在大多数情感语音数据库都采用了另一种替代方式，聘请一些专业的演员在录音室中去演绎预定的情感语音<sup>[58]</sup>。尽管有一些学者认为这样得到的语音情感表现过于夸张，和实际的自然语音不一致，但是这并不影响用这种数据库来探索声学表现和情感之间的相关性。

录制时如何唤醒情感？在录制情感语音数据库时，首先需要做的就是唤醒说话人的情感，通常有三种唤醒方式。第一种就是让说话人根据规定情感进行表演<sup>[4]</sup>，但这种方式得到情感表现过于夸张，和自然语音中的表现不一致。第二种就是将说话人置于某些特定的环境下来激起对特定的情感反应<sup>[59]</sup>，例如通过一些诱发式

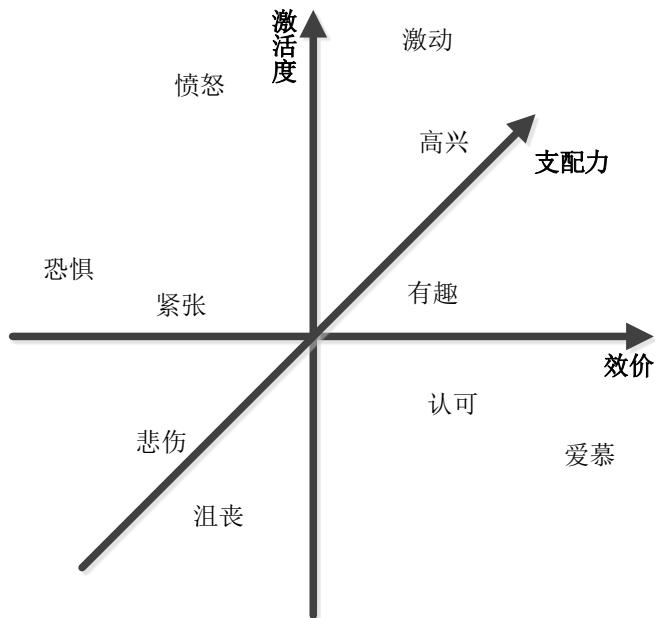


图 2.2 三维情感空间模型（激活度-效价-支配力）

的交谈，或者一些交互式的游戏。第三种就是让标注者从生活中录制的自然语音去标注出有情感的句子<sup>[60]</sup>，这种语音最为真实，但是标注成本太高，需要大量的人力劳动。

不同情感的数据量是否平衡？由于在日常生活中，不同的情感触发的几率并不相同，所以会导致包含不同情感的语音数量也不同，例如中性语音是日常生活中出现最多的。这种分布不平衡的数据库会导致在训练分类器时出现偏置，使得分类器更趋向于预测为数据量多的那种情感。有一些数据库为了保持分布平衡会保证不同情感的句子数量基本一致<sup>[61]</sup>。但一些研究者<sup>[62,63]</sup>认为这种分布正体现了实际应用场景的情感出现概率，所以应该通过调整模型来包含这种信息。

情感语音是否应该保证说话人以及说话内容无关？由于不同的人表达情感的方式不相同，如果数据库中只包含个别人的语音就会导致模型不够强健，无法识别其他人的语音。应该保证尽可能多的说话人。还有就是语音中的语言学信息通常和情感都是强相关的，在录制数据库时是否应该排除掉语言学信息的影响。现在大多数研究者的观点是对于提前准备台词的表演型数据库，由于情感触发和文本是相关联的，所以并不适合用于语音情感识别。

### 2.3.2 常用的情感语音数据库

由于大多数的情感语音数据库都不是公开的，所以只有很少的基准数据库可以被研究者们共享。但由于情感语音数据的录制没有标准的规范，所以导致不同数据库的录制方式各不相同，下面的表格列举了一些常用的情感语音数据库。

表 2.2 常用的情感语音数据库

数据库名	语言	大小	来源
LDC <sup>[64]</sup>	英语	7人×15种情感×10个句子	专业演员
柏林情感语音数据库 <sup>[61]</sup>	德语	10人×7种情感×10个句子	专业演员
丹麦情感语音数据库 <sup>[65]</sup>	丹麦语	4人×5种情感	非专业演员
Natural <sup>[63]</sup>	普通话	11人×2种情感	呼叫中心
ESMBS <sup>[16]</sup>	普通话	12人×6种情感	非专业演员
INTERFACE <sup>[66]</sup>	英语, 斯洛文尼亚语, 西班牙语, 法语	635个句子	专业演员
KISMET <sup>[24]</sup>	美式英语	3人×5种情感	非专业演员
BabyEars <sup>[25]</sup>	英语	12人×3种情感	父亲和母亲
SUSAS <sup>[62]</sup>	英语	16000个句子	压力下的模仿
MPEG-4 <sup>[67]</sup>	英语	2440个句子	美国电影
北航情感语音数据库 <sup>[68]</sup>	普通话	7人×5种情感×20个句子	非专业演员
FERMUS III <sup>[69]</sup>	德语, 英语	13人×7种情感	诱发环境
KES <sup>[70]</sup>	韩语	5400个句子	非专业演员
CLDC <sup>[71]</sup>	汉语	1200个句子	非专业演员
Pereira <sup>[72]</sup>	英语	2人×5种情感×8个句子	非专业演员
IEMODB <sup>[53]</sup>	英语	10人×9种情感	专业演员

这里主要介绍下 IEMODB<sup>[53]</sup> 这个情感语音数据库，因为本文的研究工作主要是以这个数据库作为实验基础的。IEMODB 主要被设计用于多模态情感表现研究，

它包括肢体动作，音频和视频，一共有 5 个部分，每个部分包括 10 个主题，总共有接近 12 个小时的数据。每一个部分包含一个不同的对话场景，会有一个男演员和一个女演员分别表演规定好的剧本，以及在一个对话中诱发情感。至少三个标记员对同一句话标记情感类别，包括高兴，悲伤，中性，愤怒，惊讶，激动，沮丧，厌恶，恐惧这些情感标签。这个数据库被许多的研究工作采用，因此可以用来与其他研究工作的实验结果作对比。

## 2.4 声学特征的抽取

语音情感识别中一个重要的问题就是抽取与情感相关的声学特征，因为这些特征作为模型的输入会直接影响到最终的分类效果。下面将会对特征的选择和抽取方式做出介绍。

### 2.4.1 人工选择情感相关的声学特征

在大多数研究中，声学特征都是通过以往的一些经验选择或者设计出来的。用于语音情感识别的声学特征大致可以分为三类，分别为韵律学相关的特征，谱相关的特征以及声音质量相关的特征，如图2.3所示。

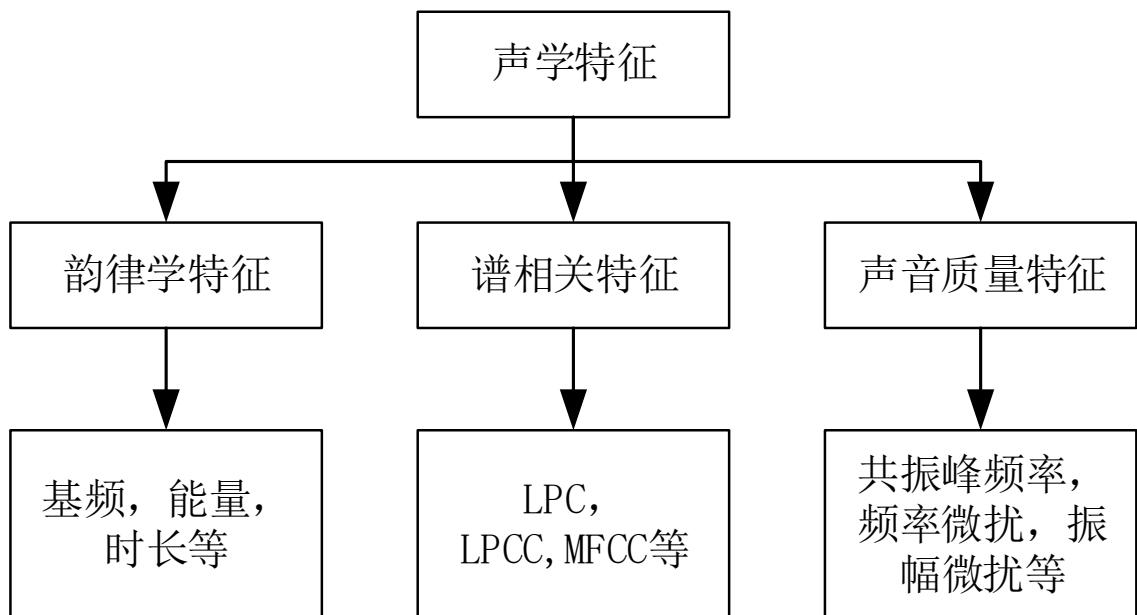


图 2.3 不同类型的声学特征

韵律是指人说话时的节奏，轻重，快慢和音高等方面的变化，它与语音中携带的语言学信息并没有太大的关联，但却决定着一句话给听众的感觉，因此又被称为“超音段特征”或“辅助语言学特征”。这种韵律学相关的特征被广泛的应用在语音情感识别领域，主要包括基频，时长，能量，共振峰等。根据 Williams 和 Stevens 的研究<sup>[73]</sup>，语音情感的激活度会显著的影响频谱上的能量分布，基频的大小以及停顿的时长，其他一些研究<sup>[74,75]</sup> 也证明了这一结论。此外，有研究证明这些特征也和基本情感类别有着很强的关联，例如 Murray 和 Arnott 的研究<sup>[5]</sup> 证明快的说话速率与愤怒是相关联的，但也有些研究<sup>[76,77]</sup> 表明部分情感的韵律学比较相似，例如愤怒，恐惧，高兴和惊讶都有相似的基频。

谱相关的特征被认为与声道对语音信号的调制相关联，这类特征之前一直被语音识别广泛的应用，但现在一些研究证明这类特征在情感识别中也发挥很大的作用，例如线性预测系数<sup>[76]</sup>，线性预测倒谱系数<sup>[6]</sup> 以及梅尔频率倒谱系数<sup>[11,78]</sup>。Nwe 等人<sup>[16]</sup> 发现语音信号不同频段的能量分布和情感类别有着相关性，例如高兴地语音通常在高频段有着较高的能量，而悲伤的语音在高频段的能量却相对较低。

声音质量特征是人对声音的一种主观评价，主要用于衡量声音的流利和清晰程度。当人在情绪比较激动的时候，通常会出现哽咽，颤音，喘息之类的反应，这会导致声音质量发生变化。因此，研究者<sup>[79]</sup> 认为声音质量特征也可以反映情感的变化。声音质量特征包括共振峰频率及其带宽、频率微扰、振幅微扰、声门参数等。Lugger 等人<sup>[36,80,81]</sup> 通过使用共振峰频率和带宽作为特征取得了很不错的效果。Li 等人<sup>[82]</sup> 也采用梅尔频率倒谱系数加频率微扰和振幅微扰，取得超过只使用梅尔频率倒谱系数的效果。

#### 2.4.2 选择算法筛选情感相关的声学特征

声学特征有许多种，选择与情感相关的特征除了依靠人工挑选以外，还可以通过一些特征选择算法来自动选出相关的特征。假设我们有一个很大的特征集合，特征选择算所要做的就是规定一种指标，例如熵增益<sup>[83]</sup> 或者识别准确率<sup>[84]</sup>，然后通过特征的各种组合来选取那些指标最好的特征子集。这样既可以减少输入特征的数量，降低计算量，又可以去除无效特征的干扰，大致流程如图2.4所示。

特征选择算法有许多，例如序列浮动前向选择算法 (Sequential Floating Forward Selection,SFFS)<sup>[85]</sup> 通过迭代的方法选择出接近最优的特征子集，还有遗传算法 (Genetic Algorithm, GA)<sup>[86]</sup> 是一种模拟生物进化的算法，它通过不断地繁殖和变异来筛选出最优的特征子集。除了特征选择算法以外，还有一些特征空间转换的算法也可以降低输入特征的维度，例如主成分分析 (Principal Component Analysis,

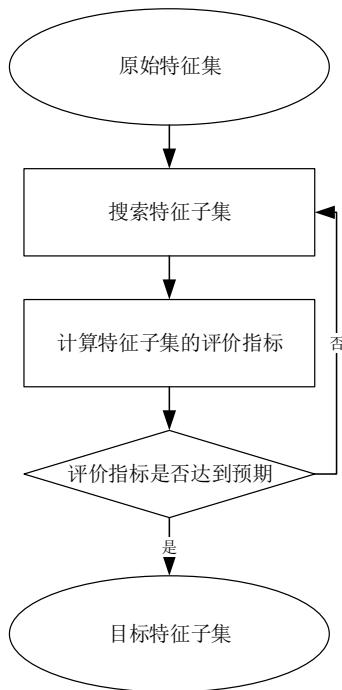


图 2.4 特征选择流程图

PCA)<sup>[87]</sup> 和线性判别分析 (Linear Discriminant Analysis, LDA)<sup>[88]</sup>，它们可以通过矩阵运算将高维特征向量转换到低维特征向量，从而减少计算量。

#### 2.4.3 深度神经网络抽取情感相关的声学特征

在近几年，深度学习方法和工具被引入了语音信号处理领域<sup>[40-45]</sup>，研究者发现采用深度神经网络从原始语音信号来提取特征，可以取得比人工定义的声学特征更好的结果，同时这也衍生出了端到端语音情感识别系统。现在有很多的神经网络结构都被用于特征抽取，例如最早的工作是 Jaitly 等人<sup>[46]</sup> 通过受限玻尔兹曼机从原始语音信号上得到一种有利于语音识别的中间表示。Bhargava 等人<sup>[47]</sup> 则是通过堆叠的全连接神经网络从原始语音信号得到瓶颈特征，并且取得了和使用梅尔频率倒谱系数相近的效果。George 等人<sup>[51]</sup> 提出一种使用 CNN 从原始语音信号提取特征，然后通过 LSTM-RNN 捕获输入序列的时序信息并最终输出不同情感的后验概率，并且他们发现 LSTM-RNN 不同节点的输出和一些声学特征有很强的相关性。Satt 等人<sup>[52]</sup> 也采用了相似的神经网络结构，但不同的是他们从语谱图上抽取特征而非原始语音信号。他们认为在语谱图上可以更方便的进行去噪的操作，并且他们在公开情感语音数据集 IEMOCAP 上取得了超过之前最好结果 (The State of Art) 的准确率。

## 2.5 情感分类模型的构建

在抽取声学特征之后，需要构建分类模型来判别特征向量所属的情感类别。事实上，目前情感语音识别的大多数研究都是关注在这一步骤上，下面我们将介绍情感分类模型的相关技术，主要包括传统的分类模型和深度学习的分类模型。

### 2.5.1 基于传统机器学习的情感分类模型

许多传统机器学习中的分类算法已经被运用在语音情感识别任务上，例如 HMM, GMM, SVM 等。目前并没有公认的最适合语音情感识别的分类器，每一种分类器都有各自的优缺点，下面我们将分别介绍几种常用的分类器模型。

HMM 被广泛地应用在语音识别领域，例如孤立词的识别和端点检测。这是因为它和语音信号的产生机制十分相似<sup>[13]</sup>。HMM 是一个包含一阶马尔科夫链的双随机过程，分别包含隐藏的转移状态和可观测的输出，其中隐藏状态是用来建模语音信号的时序信息。在数学上，为了通过 HMM 给一个可观测序列  $x_1, \dots, x_T$  建模，我们假设一个马尔科夫链可以用于生成观测序列，让  $K$  代表状态的数量， $\pi_i, i = 1, \dots, K$  代表不同状态的初始概率， $a_{ij}, i = 1, \dots, K, j = 1, \dots, K$  代表从状态  $i$  到状态  $j$  的转移概率。通常 HMM 的参数都是通过最大似然的方法来估计。假设实际的状态序列是  $s_1, \dots, s_T$ ，观测序列的似然度可以通过下面的公式给出：

$$\begin{aligned} p(\mathbf{x}_1, s_1, \dots, \mathbf{x}_T, s_T) &= \pi_{s_1} b_{s_1}(\mathbf{x}_1) a_{s_1, s_2} b_{s_2}(\mathbf{x}_2) \dots a_{s_{T-1}, s_T} b_{s_T}(\mathbf{x}_T) \\ &= \pi_{s_1} b_{s_1}(\mathbf{x}_1) \prod_{t=2}^T a_{s_{t-1}, s_t} b_{s_t}(\mathbf{x}_t) \end{aligned} \quad (2-1)$$

其中， $b_i(\mathbf{x}_t) \equiv P(\mathbf{x}|s_t = i)$  是第  $i$  个状态的输出概率。它既可以是离散的概率分布，也可以是连续的概率密度。因为真实的状态序列并不知道，所以在给定输出序列时，我们必须对所有可能的状态序列的似然度求和，如下面的公式：

$$p(\mathbf{x}_1, \dots, \mathbf{x}_T) = \sum_{s_1, \dots, s_T} \pi_{s_1} b_{s_1}(\mathbf{x}_1) \prod_{t=2}^T a_{s_{t-1}, s_t} b_{s_t}(\mathbf{x}_t) \quad (2-2)$$

幸运的是，一种计算似然度的高效算法已经提出，可以将时间复杂度降低至  $O(KT)$ 。在训练阶段，HMM 的参数通过最大化公式 2-2 的似然度来获得，这通常可以使用期望最大化 (Expectation Maximization, EM) 算法<sup>[89]</sup> 来实现。在语音识别中，HMM 的结构通常是从左到右的，因为这种结构符合语音信号的时序特性。但在语音情感识别中，除了使用从左到右结构以外，全连接的结构也会被使用，因为情感信息

可能只集中在某一个小小的时间段内，而不是所有的时间段都是均匀的。HMM 已经被许多语音情感情识别的研究所采用，在 Nwe 的工作<sup>[16]</sup> 中，一个基于 HMM 的语音情感情识别系统用于区分 6 种基本情感，模型为不同的情感和不同的说话人分别构建了一个四状态的全连接 HMM。在 Lee 的工作<sup>[18]</sup> 中，两个不同的 HMM 模型被提出，一种是普通的 HMM 加 GMM 的模型，另一种则是和语音识别一样先构建对音素的 HMM 模型，然后构建音素序列到情感类别的映射模型。作者表示采用对音素建模的方式比普通的方式可以取得更好的效果。

GMM 是一种采用多变量高斯分布的概率模型，它可以被考虑为一种只包含一个状态的 HMM<sup>[19,20]</sup>。GMM 的训练和测试过程相比于一般的连续 HMM 更为简单，这也使得它被广泛的使用在语音情感情识别，但是 GMM 无法建模语音信号的时序信息。在 GMM 模型构建中最困难的就是决定需要多少个高斯变量，最简单的方法就是人工尝试设置不同的数量，然后看模型的效果。此外，期望最大化算法可以被用来自动调整高斯变量的数量和模型的参数。GMM 相关的工作也有许多，在 Breazeal 的工作<sup>[24]</sup> 中，一个 GMM 分类器被用在 KISMET 情感情语音数据库上，获得了 77.87% 的平均准确率，后面又采用分层决策的策略取得了 81.94% 的平均准确率。GMM 也被用在其他的情感情语音数据库中，例如 BabyEars 情感情语音数据库<sup>[25]</sup>，模型尝试了 1-100 的高斯变量数量，最终在数量为 10 的时候取得了最好的结果，达到平均准确率为 75%。一个相似的结果在 FERMUS III 数据库上也被获得，16 个变量的 GMM 被用来为每种情感建模，平均准确率达到了 74.83%。

SVM 是一种非常流行的分类算法<sup>[26]</sup>，它在许多的模式识别任务中均取得了很不错的效果。SVM 模型主要是利用核函数将在低维特征空间线性不可分的向量映射到高维空间，使得数据可以被线性分类器划分。相比于 HMM 和 GMM，SVM 可以得到全局最优的分类边界<sup>[90]</sup>，但是对于不可分的数据，它又不得不采用一些启发式的方法。事实上，并没有系统性的方法可以用来选择核函数，因此，转换后的数据可分性是无法保证的。在大多数模式识别任务，包括语音情感情识别，并不建议找到训练数据的最佳分类面，因为这可能会导致过拟合。有许多语音情感情识别的研究都在使用 SVM<sup>[17,18,27,28]</sup>，它们都取得了相似的结果。其中一个工作将 SVM 通过三种不同的策略从二分类器转换为多分类器。第一种策略是将多分类任务转换为多个二分类任务，每个二分类任务将一个情感看作一类，其他所有情感看作一类，所有二分类器的结果中输出最大的情感类别代表最终的识别结果。第二种策略是将所有二分类器的输出传递给一个三层感知机，让它完成最后的决策。第三种方法是采用分层决策的策略，在不同的阶段决策不同的情感。这三种模型在 FERMUS III 数据库上做测试，分别得到了 76.12%，75.45% 和 81.29% 的分类

准确率。

除了采用单一的分类模型以外，多分类器混合模型也被用于语音情感识别，有三种不同的方法来组合不同的分类器<sup>[35,36]</sup>。第一种是分层判决的方法，每个分类器都被放置在一棵决策树的各个节点，输入从根节点出发，不断向下探索，最终到达叶子节点时会被划分到唯一的情感类别。在 Lugger 的工作<sup>[81]</sup> 中，分层的模型在柏林情感语音数据库中测试，其中采用了考虑心理学情感属性的二阶段和三阶段分层分类系统。二阶段的方法可以达到 83.5% 的分类准确率，三阶段的方法可以达到 88.8% 的分类准确率。第二种是顺序串行的判决方法，就是依次采用不同的分类器对数据进行分类，当前的分类器会影响下一个分类器的结果。第三种是多分类器并行的判决方法，就是同时训练多个分类器，然后将所有分类器的结果进行决策融合得到最终的分类结果。

### 2.5.2 基于深度学习的情感分类模型

近几年来，深度学习的模型开始变的越来越流行，这种方法也在语音情感识别中也取得了很不错的成绩。相比于传统的机器学习模型，深度学习模型可以对更复杂的非线性映射关系进行建模。已经有许多的深度神经网络结构被提出，例如自编码神经网络，CNN，RNN，RBM 等。其中一部分工作仍然是采用传统的声学特征来进行建模，但近几年也开始出现直接基于原始信号的端到端的语音情感识别。

基于传统声学特征的深度学习分类模型已经有很多工作，Han 等人<sup>[40]</sup> 采用普通的深度神经网络 (Deep Neural Network, DNN) 得到不同的语音子段在不同情感上的概率分布，然后通过统计学方法的方法将所有字段的概率分布融合得到整个句子的特征表示，最后采用一种单层的神经网络结构，叫做极限学习器 (Extreme Learning Machine, ELM) 来得到最终的情感类别。在 IEMOCAP 数据库<sup>[53]</sup> 上，这种方法取得了 48.2% 的不加权准确率 (Unweighted Accuracy, UA) 和 54.3% 的加权准确率 (Weighted Accuracy, WA)。Kim 等人<sup>[91]</sup> 提出采用深度信念网络 (Deep Belief Network, DBN) 代替传统的特征选择算法，来获得更为有效的特征表示。DBN 是由多层 RBM 堆叠而成，可以采用无监督的方式训练，在 IEMOCAP 数据库<sup>[53]</sup> 上取得了 66.12% 的不加权准确率。Deng 等人<sup>[92]</sup> 采用一种自编码神经网路来进行特征的转换学习，他们先通过大量未标记的情感语音训练自编码神经网络，从而可以找到更为有效的特征表示，然后再用标记数据对自编码神经网络进行微调，在 5 个不同的情感语音数据库上进行交叉测试均取得了不错的效果。

基于原始信号的端到端的分类模型也开始有一些工作被提出，George 等人<sup>[51]</sup>

提出一种使用 CNN 从原始语音信号提取特征，然后通过 LSTM-RNN 捕获输入序列的时序信息并最终输出不同情感的后验概率。作者认为通过 CNN 和 RNN 可以将时序信息编码到特征表示中，并且他们发现 LSTM-RNN 不同节点的输出和一些传统的声学特征有很强的相关性。实验结果表明这种端到端的方式可以取得比传统声学特征更好的效果。Satt 等人<sup>[52]</sup> 也采用了相似的神经网络结构，但不同的是他们从语谱图上抽取特征而非原始语音信号。他们认为在语谱图上可以更方便的进行去噪的操作，并且他们在公开情感语音数据集 IEMOCAP<sup>[53]</sup> 上取得了超过之前最好结果 (The State of Art) 的准确率。

## 第3章 基于情感对的语音情感情识别框架

### 3.1 本章引论

特征选择作为传统语音情感情识别中一个重要的部分，已经吸引了许多研究者的关注。因为情感是人类的主观感受，想要通过声音中的线索来反应当前说话人的情感状态是一个非常具有挑战性的任务。目前大多数的研究都旨在为所有的情感类别找到一个统一的特征集合，因为在同一个特征集合上构建分类器也是通常处理多分类问题的方法。但是一些研究结果<sup>[93]</sup> 已经证明与不同的情感相关的声学特征也是不同的，也就是说为所有的情感选择相同的特征集合并不是一个很好的方法，为特定的情感选择特定的特征集合可以取得更好的效果。

基于上述原因，我们认为给不同的情感组合选择相关程度最高的特征空间，保证这些情感在这样的特征空间上具有更高的可分性，是一个比较适合的实现方式。为此，我们提出了一种基于情感对的语音情感情识别框架。我们将任意两种不同的情感组成情感对，然后为每一个情感对选择最相关的声学特征子集。这种特征选择方式将在很大程度上缩减需要处理的问题域，因为现在我们将之前需要为多种情感选择特征的问题转换为两种情感选择特征的问题，剔除了许多无关的干扰特征。当为每一个情感对选择出对应的特征子集之后，我们将在每个特征子集上构建二分类器，这样就将原本的多分类问题转换为多个二分类问题。此时，又存在一个问题，就是最后我们期望的结果是一句话只得到一个识别结果，但现在每个二分类器都会得到一个结果，所以我们还需要再加入一个决策融合的步骤。最简单的决策融合方法就是采用投票策略，将所有二分类器的结果中出现次数最多的那个情感作为最终的识别结果。但投票的策略存在两个缺点，一是会出现票数相同的问题，二是只有和目标情感相关的那些二分类结果才有贡献，其他的二分类结果只会产生干扰。为了避免这两个缺点，我们引入情感空间模型中不同情感之间的距离信息，通过贝叶斯分类器来完成决策融合的步骤，整个系统的流程图如图3.1所示。

基于层次的语音情感情识别框架<sup>[93]</sup> 和我们的设计有着相似的思想，都是期望为不同的情感类别采用不同的特征集合，但是我们的方法效果更好，而且当情感类别发生变化时更容易扩展。此外，由于多个二分类任务之间没有依赖关系，所以更易于并行加速。

本章的剩余部分是这样安排的：首先我们将介绍情感对的定义，基于情感对的特征选择算法，和基于情感对的二分类模型；然后，我们将介绍决策融合的方

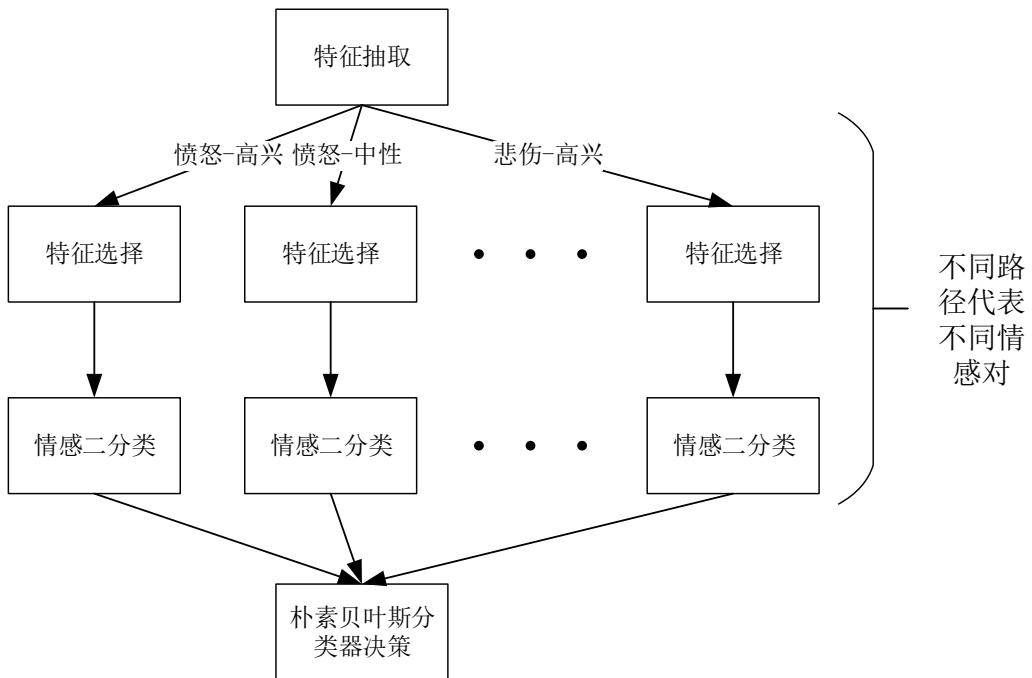


图 3.1 基于情感对的识别框架

法，包括基于投票的决策融合和基于情感空间的贝叶斯决策融合；最后，我们将和采用全局特征集合的方法，以及基于层次分类的方法进行实验对比。

## 3.2 情感对

### 3.2.1 情感对的定义

任意两个不相同的情感组合在一起就被称为情感对，这种思想是源于集成学习 (Ensemble Learning)<sup>[94]</sup>的一些观点。集成学习大致可以分为两种方式，一种被称为引导聚合 (Bootstrap Aggregating, Bagging)，这种方法利用重采样的方法从整体数据集中进行有放回的抽样得到 N 个数据集，在每个数据集上学习出一个模型，最后的预测结果利用 N 个模型的结果共同决策得到，例如随机森林 (Random Forest)<sup>[95]</sup> 就属于这种方法。总的来说，引导聚合就先训练多个简单的弱分类器，然后通过这些弱分类器组合成一个强分类器。另一种被称为提升方法 (Boosting)，这种方法是一种可以减小监督学习中偏差的机器学习算法。主要原理也是学习一系列弱分类器，并将其组合成一个强分类器，其中最具代表性的方法是 AdaBoost(Adaptive Boosting) 算法<sup>[96]</sup>，这种算法在刚开始训练时对每一个训练样例赋相同的权重，然后对训练集进行多轮迭代训练，每轮训练结束后都对预测错误的那些样例赋以较大的权重，也就是让学习算法以后更注意学错的样本，从而可以得到多个预测模

型，最终再以正比于准确率的权重将所有模型的结果组合到一起得到最后的结果。

我们提出的情感对有些类似于引导聚合的思想，我们将为不同的情感对分别选择特征和训练分类器，最后再将所有分类器的结果融合到一起。由于只需要为两种情感选择相关的特征，相比于为所有的情感选取相关的特征，这大大减少了干扰特征的引入。和情感对有相似思想的还有基于层次的语音情感情识别方法<sup>[93]</sup>，这种方法首先会根据观察设计一棵二叉决策树，在树中的不同节点分别区分不同的情感类别或情感类别组，并且每一个节点的分类器都是单独选择特征集的。情感类别组是指将多种情感分别归属到两个不同的组中，然后将这两个组看做两个类。识别过程是语音从根节点开始进行分类，然后自顶向下沿着路径上的节点依次进行分类，最后到达叶子节点后将会得到一种唯一的情感，整个分类流程如图3.2所示。但这个方法有两个缺点，第一是整个决策树的结构需要人工来设计，当情感类别或应用场景发生变化时，整个决策树就得重新设计，并不具备通用性。第二是由于整个分类过程是自顶向下的，所以会存在错误累计的问题，就是说如果上一层的分类结果是错误的，这些错分的样本就会沿着路径一直走下去，从而影响下层节点的分类效果。我们基于情感对的方法则可以解决这两个问题，首先情感对的方法不需要任何的人工介入，因为情感对只是将任意两种情感组合到一起，即使需要加入新的情感类别，仅仅只是需要生成新的情感对，之前已经训练好的二分类模型仍然可以使用，这将大大减少训练时间。其次，情感对的方法不存在错误累计问题，因为不同的情感对之间的输入都是独立的，不会相互产生影响。此外，由于不同情感对之间没有依赖关系，所以训练可以并行完成，这将进一步提升训练速度。

下面两小节我们将专门针对基于情感对的声学特征选择算法和二分类模型做出详细的阐述，这两个问题是整个识别框架中最关键的问题。

### 3.2.2 基于情感对的声学特征选择

特征选择选择算法有许多，这些算法的目的主要可以分为三个方面：第一是提升模型的预测准确率；第二是筛除无用的冗余特征，减少计算时间；第三是针对当前的问题域，对于数据提供一个更好的理解。下面将会对特征选择指标，特征选择策略以及基于情感对的特征选择实现这三个方面分别作介绍。

#### 3.2.2.1 特征评价指标

对于特征选择算法来说最重要实现选取一种指标来衡量选出的特征的好坏。假设我们有一个含有  $m$  个样例的数据集  $\{\mathbf{x}_k, y_k\}(k = 1, \dots, m)$ ，每个样例包含  $n$  个

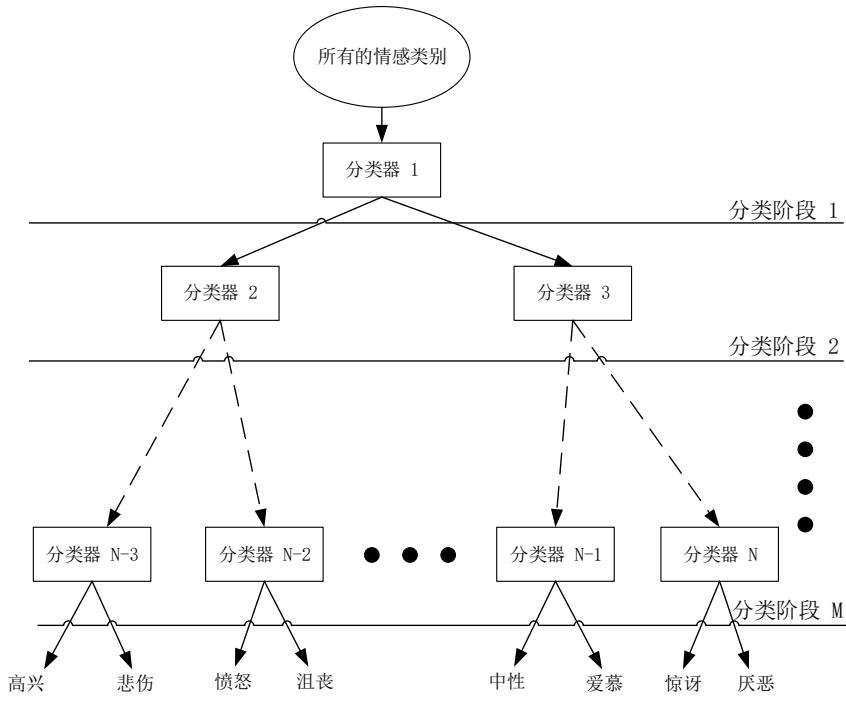


图 3.2 基于分层决策树的识别框架

输入特征  $x_{k,i}(i = 1, \dots, n)$  和一个输出值  $y_k(k = 1, \dots, m)$ , 第  $i$  个特征的评判指标通过函数  $S(i) = f(\{x_{k,i}, y_k\}(k = 1, \dots, m))$  来定义。一般来说, 我们定义  $S(i)$  越大的特征和任务目标越相关。为了方便下面的解释, 我们再引入一些其他的定义: 如果输入向量  $\mathbf{x}$  被看做来自一个潜在的多变量概率分布  $\mathbf{X}$ , 我们定义  $X_i$  代表  $\mathbf{x}$  中第  $i$  个特征的随机变量。同样,  $Y$  代表输出值  $y$  的随机变量。进一步, 我们定义  $m$  维的向量  $\mathbf{x}_i$  代表数据集中第  $i$  个特征的所有值,  $m$  维的向量  $\mathbf{y}$  代表数据集中所有样本的输出值。

第一种特征评价指标叫做相关系数<sup>[97]</sup>, 我们假设输出  $y$  是一个连续值, 则泊松相关系数被定义为:

$$\mathfrak{R}(i) = \frac{\text{cov}(X_i, Y)}{\sqrt{\text{var}(X_i)\text{var}(Y)}} \quad (3-1)$$

$\text{cov}$  代表协方差,  $\text{var}$  代表方差。通过统计方法对  $\mathfrak{R}(i)$  的估计  $R(i)$  可以表示为:

$$R(i) = \frac{\sum_{k=1}^m (x_{k,i} - \bar{x}_i)(y_k - \bar{y})}{\sqrt{\sum_{k=1}^m (x_{k,i} - \bar{x}_i)^2} \sum_{k=1}^m (y_k - \bar{y})^2} \quad (3-2)$$

符号上面的横线代表对所有的下标  $k$  求平均值, 这个系数也可以看作向量  $\mathbf{x}_i$  和向

量  $\mathbf{y}$  的余弦值。在线性回归中，评价指标通常是  $R(i)$  的平方，这样将可以去除负值，仅仅表示向量  $\mathbf{x}_i$  和向量  $\mathbf{y}$  的线性相关性。此外，相关系数  $R(i)$  仅仅可以表示特征和目标之间的线性依赖关系，如果希望可以获取非线性关系，最简单的方法就是对输入进行非线性的处理。例如取平方，开方，取对数等等。

第二种方法是采用单变量分类器<sup>[98]</sup>，这种方法是将一个特征输入一个分类器中，然后将分类器的预测能力作为衡量指标。通常用来衡量分类器的预测能力的指标是错误率，例如对于二分类问题，错误的识别为正例的比率 (False Positive Rate, FPR) 和错误识别为负例的比率 (False Negative Rate, FNR) 都可以被定义为衡量标准，但通常为了平衡两种错误率，会选择 ROC 曲线 (Receiver Operating Characteristic Curve) 下区域的面积，也就是 AUC(Area Under Curve) 作为衡量的指标。

第三种方法是利用信息学理论中的一些指标，最常使用的是变量和目标之间的互信息量<sup>[98,99]</sup>，定义如下：

$$I(i) = \int_{x_i} \int_y p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)p(y)} dy dx_i \quad (3-3)$$

其中  $p(x_i)$  和  $p(y)$  是  $x_i$  和  $y$  的概率密度， $p(x_i, y)$  是联合概率密度， $I(i)$  是用来衡量变量  $x_i$  的概率密度和变量  $y$  的概率密度之间的相关性。连续变量的概率密度  $p(x_i)$ ， $p(y)$  和  $p(x_i, y)$  都是未知的，并且很难从数据集中估计出来。通常在估计连续变量的概率密度时，会先假设变量服从某种已知的概率分布，例如高斯分布，然后通过训练数据中变量的统计值来估计这种分布的参数。相对而言，离散变量的概率分布是更容易估计的，因为积分可以通过求和来替代，计算公式如下：

$$I(i) = \sum_{x_i} \sum_y P(X = x_i, Y = y) \log \frac{P(X = x_i, Y = y)}{P(X = x_i)P(Y = y)} \quad (3-4)$$

公式中的概率可以通过统计不同值出现的频数。例如，在一个三分类的问题中，输入一共有 4 个特征， $P(Y = y)$  代表类别的先验概率 (3 种可能)， $P(X = x_i)$  代表输入特征的分布 (4 种可能)， $P(X = x_i, Y = y)$  代表联合概率 (12 种可能)，但是当类别和特征的数量增多时，这种估计也将变得更加困难。

### 3.2.2.2 特征选择策略

上一章我们列举了衡量特征的指标，但是上面的指标主要是针对于单个特征来说的。通常我们选择特征时会选择一个特征子集，并不是说特征子集中所有单个特征的指标最好就代表特征子集是最好的。因为一些研究已经证明，不同特征

组合在一起时，相互之间会产生影响。一些评价指标比较低的特征可能刚好补充了其他特征缺失的那一部分信息，一些评价高的特征也有可能和其他特征所包含的信息有重复，所以我们需要通过一些方法来保证能够选择到好的特征子集。特征子集选择的策略大致可以分为两类，一类叫作打包 (Wrappers)<sup>[100]</sup>，它是通过预测模型在采用不同特征子集时的预测能力来评价的；另一类叫作过滤 (Filters)<sup>[101]</sup>，是通过一些预处理步骤得到特征子集，与预测模型无关，下面分别介绍两类方法。

打包 (Wrappers) 是一种简单和有效的特征选择策略。在通常情况下，打包 (Wrappers) 通过预测模型的预测能力来评价特征子集的有效性，所以需要解决的有三个问题：一是如何能够遍历所有的特征子集，二是如何通过预测模型的预测能力指导遍历过程，三是应该选择哪种预测模型。当特征数量不是太多的时候，完全遍历将会被采用。但完全遍历本身是一个 NP-hard 问题，当特征数量增多时，完全遍历的计算量是难以承受的，一些启发式的遍历算法被提出来降低计算量，期望得到一个近似最佳的结果，例如局部贪心，剪枝搜索，模拟退火，遗传算法等。预测能力则可以使用评价指标中提到的错误率来衡量。预测模型的选择没有太多的理论指导，通常都是根据经验来选取。由于特征选择的过程和选择的预测模型是相关的，因此大多数人通常会将特征选择和模型训练作为一个整体来完成。

过滤 (Filters) 是另一种特征选择的策略，这种策略不需要预测模型的参与，而是利用一些信息学理论中的指标，仅仅在数据预处理的阶段就能够完成。相比于打包 (Wrappers) 的策略，一些研究者认为过滤 (Filters) 的策略有两个主要的优点：一是由于不涉及模型训练，所以处理速度更快；二是这种方法并不针对于某种特定的模型，所以选出的特征更有普适性，在所有的预测模型中都可以使用，不易产生过拟合。这种算法主要利用互信息量这种评价指标，相关的算法比较少，比如马尔科夫毯 (Markov Blanket)。

除了上面提到的特征子集选择以外，还有一些特征空间压缩的方法具有相同的效果。特征空间压缩是指将原始的特征空间通过某种函数关系映射到更低维的特征空间，进而提升效果并减少计算量。此外，构建新的特征空间可以帮助我们更好地理解所处理的问题。特征空间压缩的算法有很多，包括聚类方法，线性转换，小波变换，卷积核等等，下面我们主要介绍聚类和矩阵变换这两类方法。

聚类方法被广泛地用在特征重构上面，这种方法的思想是将一组相似的特征用他们的聚类中心所替代，变成一个新的特征，最流行的算法包括 K 均值 (K-means) 和分层聚类<sup>[102]</sup>。聚类通常是一种无监督的算法，但也可以引入一些监督信息来得到更加有效的特征。假设  $\tilde{X}$  是一个代表重构特征的随机变量， $X$  代表原有的特征， $Y$  代表预测目标。监督式算法的目的是在保证  $\tilde{X}$  和  $Y$  的互信息量  $I(\tilde{X}, Y)$  同时，最

小化  $X$  和  $\tilde{X}$  的互信息量  $I(\tilde{X}, X)$ 。这可以通过引入拉格朗日算子  $\lambda$  来构建全局目标函数：

$$J = I(\tilde{X}, X) - \lambda I(\tilde{X}, Y) \quad (3-5)$$

这使得在搜索最大可能的压缩解的同时又可以保证对目标的潜在信息。

矩阵变换是另一种特征空间压缩的方法，常用方法有 PCA<sup>[103]</sup>, SVD<sup>[104]</sup>, LDA<sup>[105]</sup> 等。这里我们简单解释一下这三个算法的原理，PCA 的目的是保证新的特征之间的方差最大，这样可以最大程度保留原始特征的信息。SVD 和 PCA 类似，目标是构成一组由原始特征通过线性组合得到的新特征，并且最大可能保证新的特征保留原始特征的信息，但 SVD 可以应用在行数和列数不等的矩阵。PCA 和 SVD 都属于无监督的算法，但 LDA 是一种有监督的算法，它的目标函数是在保证类别内样本的方差尽可能小，类别间样本中心点的距离尽可能大。

### 3.2.2.3 基于情感对的特征选择实现

在对情感对选择特征子集的时候，我们的目的是从一个大的声学特征集合中选择出一个最能够区分当前情感对中两种情感的特征子集。这里我们采用了打包(Wrappers)的方法，就是将分类器的识别率作为我们筛选特征的指标。首先需要一种遍历算法得到所有可能的特征子集，如果采用全局遍历的话需要耗费太多的时间，所以我们采用了一种启发式的遍历算法：序列浮动前向选择算法(Sequential Floating Forward Selection,SFFS)<sup>[85]</sup>。

SFFS 的算法原理是，假设我们选择的特征子集为  $S$ ，初始时  $S$  为空，每轮迭代从总的特征集合  $W$  中选出一个子集  $A$ ，使得  $A$  加入  $S$  后的评价函数  $J(S)$  达到最优，然后再从  $S$  中选择一个子集  $B$ ，使得  $S$  剔除  $B$  后的评价函数  $J(S)$  达到最优。经过多轮迭代，评价函数  $J(S)$  达到我们设定的阈值后算法停止。SFFS 被广泛的应用在许多的模式识别任务中，可以在可接受的时间范围内选择出近似最优的特征子集。我们将会为每个情感对都运行 SFFS 算法得到相关的特征子集，整个特征选择的流程图如图3.3所示：

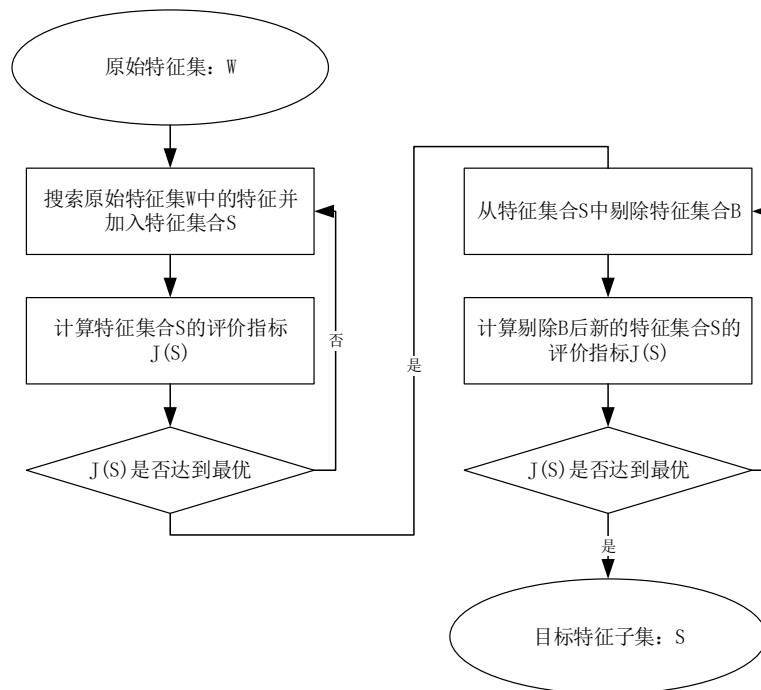


图 3.3 SFFS 特征选择算法流程图

### 3.2.3 基于情感对的二分类模型

当我们得到每个情感对相关的声学特征集后，下面的任务就是为每个情感对训练二分类器。二分类器的算法有很多种，在论文中我们采用了两种简单的分类器模型：支持向量机 (Support Vector Machine, SVM)<sup>[26]</sup> 和贝叶斯逻辑回归 (Bayesian Logistic Regression, BLR)<sup>[106]</sup>。

SVM 是一种二分类模型，它的基本原理是在特征空间中寻找使数据点的间隔最大化的分类超平面。理想情况下，所有的训练样本都是线性可分的，SVM 则可以找到一个完美的分类面将两类数据区分。但大多数情况下，训练样本并不是线性可分的，所以需要其他的一些方法来保证 SVM 能够正常工作。一般有两种方法：软间隔 (Soft Margin) 和核函数 (Kernel Function)。软间隔通过在目标函数中引入松弛变量，从而解决由于个别离群点而导致得到的分类超平面偏移到不好的位置。核函数则是通过将原始数据点映射到给高维的特征空间，使得在低维空间线性不可分的数据点在高维空间变得线性可分。

BLR 是一种通过假设变量服从某种分布的逻辑回归 (Logistic Regression) 模型。首先，让我们先回顾下普通的逻辑回归模型，假设我们有特征向量  $\phi$ ，两个类别  $C_1$  和  $C_2$ ，我们需要建模后验概率  $P(C_i|\phi)(i = 1, 2)$ ，逻辑回归模型采用下面的公式来

建模：

$$P(C_i|\phi) = \sigma(w^T \phi) \quad (3-6)$$

其中  $w$  代表我们需要学习的参数， $\sigma$  代表 sigmoid 函数。通过最大似然的算法就可以学习到参数的值。普通的逻辑回归模型都不假设后验概率服从某种分布，这也导致在拟合参数时和训练数据是强相关的，容易出现过拟合的情况。而贝叶斯方法通常都会假设后验概率的分布情况，训练数据通常都是用来估计概率分布的参数，然后再用概率分布去估计类别的概率，这样可以减少过拟合的情况，但是如果数据不服从假设的分布，也会导致训练结果变差。BLR 的目的就是将贝叶斯方法的概率分布假设引入逻辑回归的参数估计中，具体来说就是在模型训练时不再精确地估计参数  $w$ ，而是假设参数  $w$  服从某种分布，然后转而去估计这种分布的参数，这可以通过拉普拉斯近似的方法来实现，具体的细节可以参考<sup>[106]</sup>。通过这种方式就可以有效的避免过拟合的出现，尤其是在训练数据比较少的情况下。

### 3.3 决策融合

在前面一章，我们已经为所有的情感对选择了相关的特征集，并且得到了二分类结果，但我们的目标是只得到唯一的情感类别，所有还需要一个步骤来将这些情感对的结果进行汇总，也就是决策融合<sup>[107]</sup>。决策融合的算法最早出现在集成学习 (Ensemble Learning) 领域，因为集成学习通常需要训练多个模型，然后通过决策融合将所有模型的结果整合到一起。决策融合大致可以分为两种：训练型和非训练型。训练型的决策融合就是需要通过训练来建立所有模型的结果和最终结果之间的映射关系，非训练型的决策融合则是通过某些代数规则来将所有模型的结果映射到最终结果。下面会介绍本文中采用的两种决策融合的方法：基于投票的决策融合和基于情感空间的贝叶斯决策融合。

#### 3.3.1 基于投票的决策融合

每一个情感对包含两种情感，二分类器可以得到这两种情感中的一种。假设我们需要识别  $M$  种情感，则一共可以组成  $C_M^2 = \frac{M \times (M-1)}{2}$  个情感对，同样我们也会得到这么多的二分类结果。在所有的二分类结果中，每一种情感出现的次数最多为  $M - 1$  次。基于投票的决策融合就是将所有二分类结果中出现次数最多的那个情感判定为最终的识别结果，因为通常出现次数多代表语音中包含这种情感的信息最多，这属于一种无需训练的决策融合方法。下面是算法描述：

---

**算法 3.1 投票决策算法**

---

**输入:** $M$ : 情感类别的数量 $E = \{e_i, i = 1, 2, \dots, M\}$ : 情感类别的集合 $R = \{r_{e_i e_j} | e_i \neq e_j; r_{e_i e_j}, e_i, e_j \in E\}$ : 情感对的二分类结果**输出:** $f$ : 最终识别出的情感类别

- 1: 计算  $R$  中不同情感类别出现的次数  $N_e = \{n_{e_i} | e_i \in R\}$
  - 2: 选出  $N_e$  中次数最多的情感类别, 构成候选情感类别集合  $C_{max} = \{c_k | c_k \in E; k = 1, 2, \dots, K\}$
  - 3:  $f := c_1$
  - 4: **if**  $K > 1$  **then**
  - 5:     **for**  $k = 2$  to  $K$  **do**
  - 6:          $f := r_{fc_k}$
  - 7: **return**  $f$
- 

投票决策是选取票数最多的情感类别作为最终的结果, 但可能会存在出现多个相同最多票数的情感类别的情况。在本文采用投票决策算法中, 会将所有最大票数的情感类别放入一个候选集中, 然后依次比较该候选集中两个情感所在情感对的二分类结果, 保留胜出的那个情感类别, 最后将会只剩下唯一的一种情感类别。下面我们将作出形式化的证明:

**命题 3.1:** 投票决策算法在二分类结果正确的情况下一定可以得到正确的情感类别

**证明** 采用和算法 3.1 相同的符号表示, 假设  $e_i$  是目标情感, 则可以得到下面的证明过程:

$$\begin{aligned} R \text{ is correct} &\Rightarrow n_{e_i} = M - 1 \\ &\Rightarrow n_{e_j} < M - 1, e_j \in E - \{e_i\} \\ &\Rightarrow f = e_i \end{aligned}$$
□

从证明中可以看出当二分类结果是正确的时侯, 投票决策算法一定可以得到正确的情感类别, 但通常我们无法保证所有二分类器的结果都是正确的, 这会导

致一些情感的票数相同。尽管在算法 3.1 中我们采取了一些策略保证最后可以得到唯一的情感，但是如果同票的几种情感的二分类结果间的胜出关系出现闭环，我们的投票策略同样无法保证得到的一定是最好的识别结果。假设有三种同票的情感  $e_1$ ,  $e_2$  和  $e_3$ ,  $e_1 > e_2$  代表在  $e_1$  和  $e_2$  组成的情感对二分类结果中， $e_1$  的概率更高。如果出现  $e_1 > e_2$ ,  $e_2 > e_3$ ,  $e_3 > e_1$  这种情况，投票策略将无法做出有效的判决。此外，投票策略在判断最终的情感类别时是通过二分类器的结果决策，也就是说只有包含目标情感的那些二分类器才对最终的结果有好的影响，而其他二分类器的结果都不会对识别出正确的情感有好的影响。

### 3.3.2 基于情感空间的贝叶斯决策融合

鉴于上一节提到的关于基于投票的决策融合方式存在的问题，我们又提出了一种新的基于情感空间模型<sup>[3]</sup> 的贝叶斯决策融合来避免这些问题。我们在2.2.2节提到过关于维度情感空间的定义，心理学上将情感通过一个笛卡尔空间坐标系来表示，其中每一个坐标轴都表示一种心理学属性，而不同的情感类别会被映射到空间中不同的位置。常用的情感空间模型有二维情感空间(激活度-效价)和三维情感空间(激活度-效价-支配力)，这里采用三维情感空间模型来描述。假设当前需要识别 4 种情感：高兴，悲伤，愤怒和中性，这些情感在维度情感空间中的分布大致如图3.4所示。

从图中我们可以看出不同的情感类别之间的距离并不相同，例如高兴和愤怒的距离明显比高兴和悲伤的距离要小，而这种情感类别之间的距离恰好反映了它们之间的相似程度。回想一下日常生活中，当我们高兴地说话和愤怒地说话时，通常语气都比较急促，声音都比较大，但是当我们悲伤地说话时，声音一般都会比较缓慢而低沉。投票决策有个问题就是不包含目标情感的情感对的二分类结果对最后的决策并没有贡献，但现在有了维度情感空间的这种信息，我们可以认为这些不包含目标情感的情感对的二分类结果对最后的决策也是有帮助的。假设目标情感是高兴，对于愤怒-悲伤这个情感对来说，二分类结果是愤怒就更加支持最终的情感类别是高兴，是悲伤的话就更不支持最终的情感类别是高兴，所以二分类结果是愤怒的可能性会更大。这样的支持关系对其他的情感也是适用的，所以如果这种信息能够被有效地运用在决策融合当中，是有助于我们更好地判断最终情感类别的。

对于这种支持关系，需要通过一种数学模型来对其进行量化表示，最为自然的就是想到贝叶斯概率模型，因为这种支持关系可以通过条件概率来建模。假设需要识别的情感类别集合为  $E = \{e_i | i = 1, 2, \dots, M\}$ ，情感对的所有二分类结果为

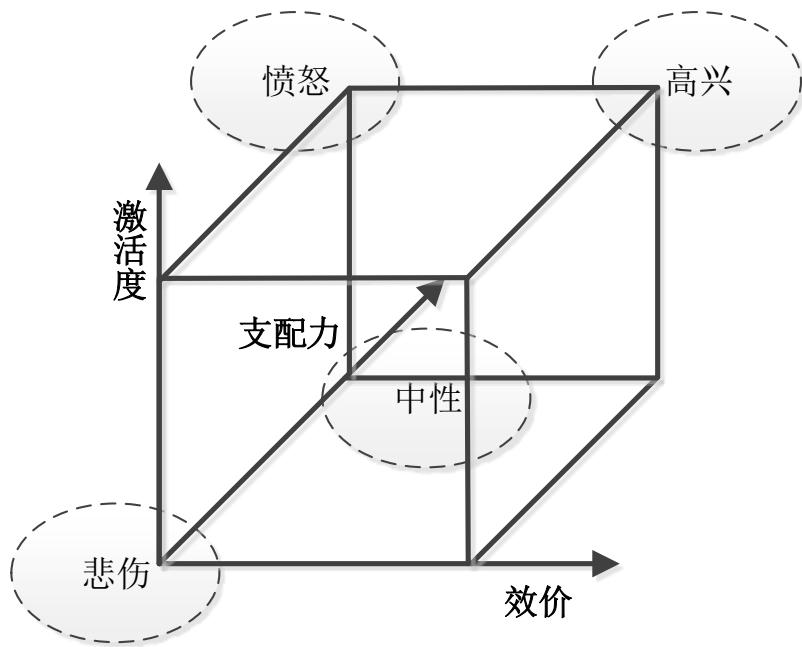


图 3.4 情感空间模型中不同情感的位置

$R = \{r_{e_i e_j} | e_i \neq e_j; r_{e_i e_j}, e_i, e_j \in E\}$ 。上面关于目标情感为高兴时，对愤怒-悲伤情感对的二分类结果影响程度的例子，通过概率的形式可以表示如下 (H: 高兴，A: 愤怒，S: 悲伤)：

$$P(r_{A_S} = A|H) > P(r_{A_S} = S|H) \quad (3-7)$$

基于贝叶斯定理，高兴和愤怒-悲伤情感对的二分类结果之间的关系可以被转换为下面的表示：

$$P(H|r_{A_S}) = \frac{P(r_{A_S}|H)P(H)}{P(r_{A_S})} \quad (3-8)$$

$$P(H|r_{A_S}) \propto P(r_{A_S}|H) \quad (3-9)$$

当  $P(H)$  和  $P(r_{A_S})$  都是先验概率时，通过公式3-8和公式3-9，我们可以推导出公式3-10：

$$P(H|r_{A_S} = A) > P(H|r_{A_S} = S) \quad (3-10)$$

为了将这种概率关系推广所有情感对二分类结果的决策融合中，我们通过朴素贝叶斯分类器来建模。目标情感  $e_i$  的后验概率可以表示如下：

$$P(e_i|R) = \frac{P(R|e_i)P(e_i)}{P(R)} \quad (3-11)$$

由于不同情感对的二分器分别实在不同的特征集上训练，所以我们可以假设这些二分类器的输出都是条件独立的。这个假设在数学上并不严谨，但这对我们最终的推导结果并没有太大的影响，因为最终的结果只是一个比例关系。基于这种假设，公式3-11可以被表示为下面的公式3-13：

$$P(e_i|R) = \frac{\prod_{r_{e_j e_k} \in R} P(r_{e_j e_k} | e_i)P(e_i)}{P(R)} \quad (3-12)$$

类似公式3-8到公式3-9的转换，我们也可以通过公式3-13推导出下面关于目标情感和情感对二分类结果之间的关系：

$$P(e_i|R) \propto \prod_{r_{e_j e_k} \in R} P(r_{e_j e_k} | e_i) \quad (3-13)$$

通过上面的概率推导，我们能够得出这样的结论：当采用朴素贝叶斯分类器作决策融合时，不同情感在维度情感空间中的距离信息能够被引入进来。这使得不仅包含目标情感的情感对的二分类结果对最终的决策有影响，而且不包含目标情感的情感对的二分类结果也可以对最终决策提供辅助信息，进而提升情感情识别的准确率。此外，基于贝叶斯的决策融合不存在投票决策中存在的同票问题。在模型训练时，我们将首先得到情感对的二分类结果，再通过朴素贝叶斯分类器建立这些二分类结果和最终目标情感的映射关系。

### 3.4 实验结果及分析

前面几节详细的介绍了我们提出的基于情感对的特征选择和不同的决策融合方法，这两部分共同构成了整个语音情感情识别系统，下面我们将通过实验来对比基于情感对的语音情感情识别框架和其他语音情感情识别框架之间的识别效果。

### 3.4.1 实验设置

#### 3.4.1.1 情感情音数据库

在本章中，所有的实验都是基于 IEMOCAP 情感情音数据<sup>[53]</sup> 库来进行的。在2.3节，我们已经对 IEMOCAP 数据库做了一个简要的介绍。这是一个人类情感交流相关的数据库，主要被设计用于多模态情感表现研究的，它包括肢体动作，音频和视频，一共有 5 个部分，每个部分包括 10 个主题，总共有接近 12 个小时的数据，语言为英语。每一个部分包含一个不同的对话场景，会有一个男演员和一个女演员分别表演规定好的剧本，以及在一个对话中诱发情感，所有的过程都是在专业的录音棚中进行。至少三个母语为英语的标记员对同一句话标记情感类别，包括高兴，悲伤，中性，愤怒，惊讶，激动，沮丧，厌恶，恐惧这些情感标签。除了离散的情感标签，标记员还需要通过 FeelTrace 软件标注每个句子在维度情感空间的心理学属性。这个数据库被许多的研究工作采用，因此可以用来与别人的实验结果作对比。在本文中，我们将只采用其中的语音数据，并且只使用那些三个标记员中至少有两个标记为同一种情感标签的句子，这样的句子表达的情感更为清晰。此外，由于数据库中不同情感标签的句子数量不相同，尤其一些情感的句子数量和其他情感相差太多，所以在我们的研究中将只针对愤怒，高兴，悲伤和中性这四种情感。因为这几种情感的语音数据相对较多，同时这也是大多数使用这个数据库的研究所采用的配置。下面的表格3.1列举了属于不同的情感的句子数量。

表 3.1 IEMOCAP 数据库<sup>[53]</sup> 中不同情感的句子数量

中性	愤怒	高兴	悲伤	总计
1708	1103	595	1084	4490

本文主要针对于说话人无关的语音情感情识别，在实验中我们将不同说话人的数据区分开。数据库由 10 个演员录制，所以分为 10 份语音数据，其中 1 个人的数据作为测试集，1 个人的数据作为验证集，剩下 8 个人的数据作为训练集，并且会采用交叉验证 (10-Fold Cross Validation)，将 10 个演员的数据依次作为测试集，最后的准确率是所有测试结果的平均值。交叉验证的方法可以在统计学意义上保证实验结果的有效性和可泛化性，使得模型不会因为数据的改变而产生较大偏差。

#### 3.4.1.2 声学特征集合

表格3.2展示了我们在实验中所采用的声学特征列表和使用的统计函数，这个特征集合被用在 Interspeech2009 语音情感情识别比赛上，因此被许多研究作为基准

特征集来使用。我们采用 OpenSmile 工具包来抽取这些声学特征，特征集合中有 16 个低级描述子 (Low Level Descriptors)，包括韵律学特征，谱相关的特征和声音质量特征。由于这些特征都是基于语音帧来抽取，所以将采用统计函数来得到整个句子的特征表示，最终对于每一个句子将会得到 384 个特征。后面的所有实验都是基于这个特征集合进行的，对于使用全局特征的模型将会输入所有的 384 维特征值，而对于选择特征子集的模型将会从这个特征集中选择出最符合当前任务的特征子集。通常会将每个特征子集中特征的数量控制在 40-60 个，这样设置是为了和之前相关的工作保持一致，从而便于比较。

表 3.2 声学特征列表及其统计函数

原始特征类型	统计函数
基频 (f0)	均值, 标准差, 峰值
均方根能量 (rms)	斜率, 最小值, 最大值
过零率 (zcr)	相对位置, 范围
信噪比 (hnr)	二次线性回归系数
梅尔频率倒谱系数 (1-12 mfcc)	线性回归均方差

在得到每个句子的特征值后，我们采用 z-normalization 对每一维特征进行正则化，具体做法是计算训练集所有中性样本中每一维特征的均值和标准差，然后将所有样本的对应特征减均值除标准差。这里我们做了一个假设：所有说话人在中性语音上的差异不是很大。通过这种正则化后，可以比较好的消除掉不同说话人的声音特点。

### 3.4.2 实验结果

实验结果主要可以分为两个部分，第一部分主要比较了基于情感对的语音情感识别框架和其他的方法的准确率，并且分析了不同情感的混淆矩阵 (Confusion Matrix) 的变化；第二部分主要展示当目标情感不同时，所有情感对的二分类结果的分布直方图，从而验证维度情感空间中不同情感的距离信息是能够反应不同情感的相似程度，并且这种相似是可以在模型输出中观测到的。

#### 3.4.2.1 准确率对比

首先我们将对比不同方法的加权准确率 (Weighted Accuracy, WA) 和不加权准确率 (Unweighted Accuracy, UA)，其中 WA 的计算方式是用测试集中正确分类的样本数量除以总的样本数量，能够反映分类器模型的整体效果，而 UA 的计算方式则

首先计算出所有情感类别的准确率，然后取平均值，这样的得到的准确率可以反映模型对于不同的情感的识别是否均衡。这两个评价指标在之前许多情感情识别相关的比赛中被使用，可以方便和其他人的工作作对比。

在下面的表3.3中，我们展示了多种语音情感情识别方法在上面的实验设置中得到的 WA 和 UA。其中，“全局特征选取”代表对于将语音情感情识别建模为多分类任务，为所有的情感选取相同的特征集合；“决策树分层特征选取”代表采用基于决策树的语音情感情识别框架，在不同树节点选取不同的特征子集；“情感对特征选取”代表采用基于情感对的语音情感情识别框架，为不同的情感对选取不同的特征子集；“+”后面是使用的分类器，“SVM”代表支持向量机，“BLR”代表贝叶斯逻辑回归；对于情感对的方法，还列出了不同的决策融合方法，“投票决策”代表采用投票的决策融合，“贝叶斯决策”代表采用考虑情感空间信息的贝叶斯决策融合。从实验结果中可以看到，基于情感对的框架在不同的分类器上都可以取得比其他框架更好的 WA 和 UA。此外，在情感对的框架中，当我们采用贝叶斯决策方法时，可以取得比采用投票方法更好的效果，这是因为引入不同情感在情感空间中的距离信息有助于在最后决策时更好的判断最终的情感类别。

表 3.3 不同方法的准确率

模型	WA	UA
全局特征选取 + SVM	52.41%	51.02%
全局特征选取 + BLR	53.47%	53.55%
决策树分层特征选取 + SVM	55.98%	58.76%
决策树分层特征选取 + BLR	56.38%	58.46%
情感对特征选取 + SVM + 投票决策	55.53%	58.16%
情感对特征选取 + BLR + 投票决策	56.15%	58.54%
情感对特征选取 + SVM + 贝叶斯决策	57.23%	62.16%
情感对特征选取 + BLR + 贝叶斯决策	<b>57.85%</b>	<b>62.54%</b>

为了进一步分析不同情感的识别率，我们将展示不同的方法得到的混淆矩阵。为了简单起见，我们将只列出当分类器为 BLR 时的混淆矩阵，因为 SVM 得到的结果和 BLR 是相似的，对于情感对的框架，我们也只展示了贝叶斯决策的结果，因

为效果更好。下面的表3.4, 表3.5和表3.6分别是三种不同的框架得到的混淆矩阵, 从中我们可以看到基于决策树的框架和基于情感对的框架在所有情感的准确率上都要优于全局选取特征的多分类框架, 这证明为不同的情感选取不同的声学特征是有助于提高语音情感识别系统效果的。此外, 相对于决策树的框架, 我们可以发现情感对的框架在愤怒, 高兴, 悲伤这三种情感的准确率上都更高, 这是因为当我们引入情感空间的距离信息之后, 不同情感之间的混淆在决策融合阶段被减轻了。此外, 中性情感的准确率并没有发生明显的变化, 这可能是因为中性情感在情感空间中和其他情感的距离没有明显的差别, 这使得贝叶斯决策并没有产生帮助。这些实验数据证明了我们提出的基于情感对识别框架在非中性情感的识别上是优于基于决策树的框架的。

表 3.4 混淆矩阵 (全局特征选取 + BLR)

预测 实际 \ 预测	中性	愤怒	高兴	悲伤
中性	<b>52.41%</b>	7.98%	16.31%	23.30%
愤怒	17.64%	<b>60.55%</b>	18.32%	3.49%
高兴	25.32%	24.58%	<b>39.28%</b>	10.82%
悲伤	30.12%	2.82%	5.10%	<b>61.96%</b>

表 3.5 混淆矩阵 (决策树分层特征选取 + BLR)

预测 实际 \ 预测	中性	愤怒	高兴	悲伤
中性	<b>54.51%</b>	6.89%	15.20%	23.40%
愤怒	16.62%	<b>65.40%</b>	15.26%	2.72%
高兴	26.13%	19.57%	<b>41.72%</b>	12.58%
悲伤	21.70%	2.22%	3.88%	<b>72.21%</b>

### 3.4.2.2 情感对二分类结果统计

尽管在识别准确率上显示出考虑情感空间信息的贝叶斯决策方式能够取得更好的结果, 但为了进一步证明情感对的二分类结果是符合我们对于情感间的距离能够反映相似度的假设, 下面的图3.5、图3.5、图3.5、图3.5将分别展示出当目标情感不同时, 所有二分类结果中不同情感的统计直方图。直方图的计算过程是找到测试集中特定目标情感的所有句子, 然后把它们输入所有情感对的二分类器, 最

表 3.6 混淆矩阵（情感对特征选取 + BLR）

实际 \ 预测	中性	愤怒	高兴	悲伤
中性	<b>53.98%</b>	6.19%	12.39%	27.43%
愤怒	15.46%	<b>68.04%</b>	12.37%	4.12%
高兴	21.94%	15.82%	<b>50.51%</b>	11.73%
悲伤	14.93%	1.49%	5.97%	<b>77.61%</b>

后统计这些二分类结果中每一种情感出现的次数，其中横轴代表不同的情感类别，纵轴代表不同情感出现的次数占所有二分类结果的比例，更高的比例表示该情感和目标情感的相似度更高。从图中可以看出，这种比例关系的确和情感空间中不同情感的距离是相关的，例如当目标情感是高兴时，愤怒的比例相对于悲伤的比例更高，而情感空间中也是愤怒和高兴地距离比悲伤和高兴的距离更近。这个统计直方图表明对于情感空间中的距离信息能够反映情感相似度的假设是正确的。

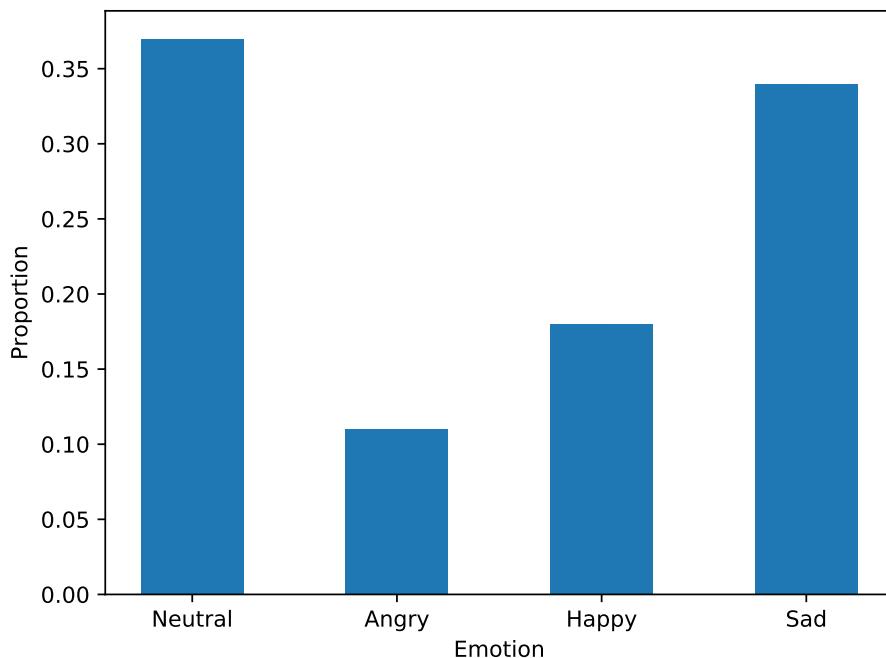


图 3.5 二分类结果中不同情感类别的分布直方图（目标情感为中性）

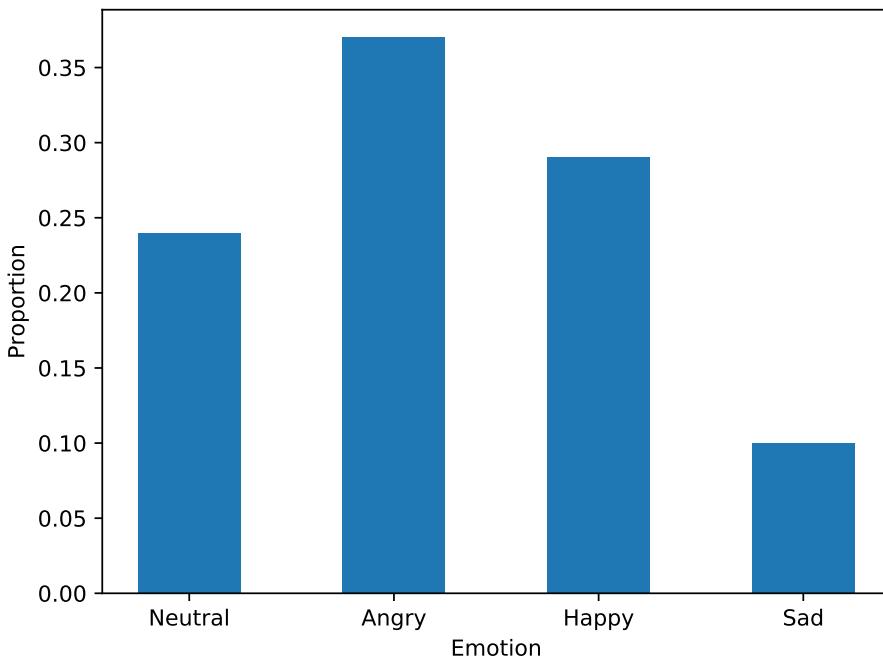


图 3.6 二分类结果中不同情感类别的分布直方图（目标情感为愤怒）

### 3.5 本章小结

本章我们介绍了基于情感对的语音情感情识别框架，通过将任意两种不同的情感组成情感对，并且为每一个情感对从一个大的声学特征集合中挑选出适合的特征子集来训练二分类器，最后再将维度情感空间中不同情感的距离信息加入到决策融合的过程中，从而避免了投票策略中出现的同票问题和无效情感对的问题。在同样的语音情感数据库 IEMOCAP 以及同样的实验设置下，相比于传统的全局选取特征的多分类框架和基于决策树的识别框架，基于情感对的识别框架加上贝叶斯决策融合可以达到更好的识别准确率，并且当情感类别发生变化时更方便构建。同时，由于情感对的二分类器训练相互独立，所以可以并行进行。

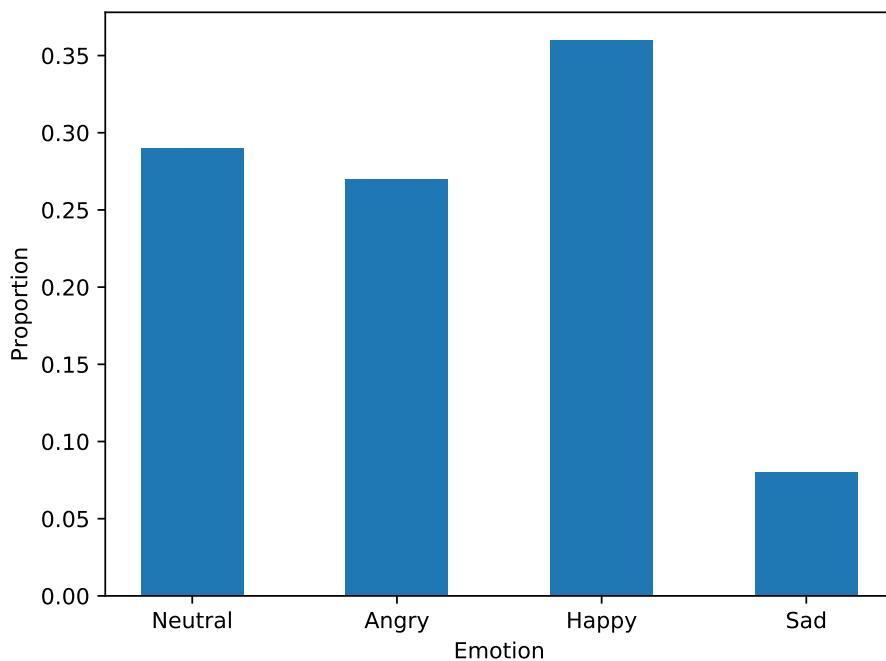


图 3.7 二分类结果中不同情感类别的分布直方图（目标情感为高兴）

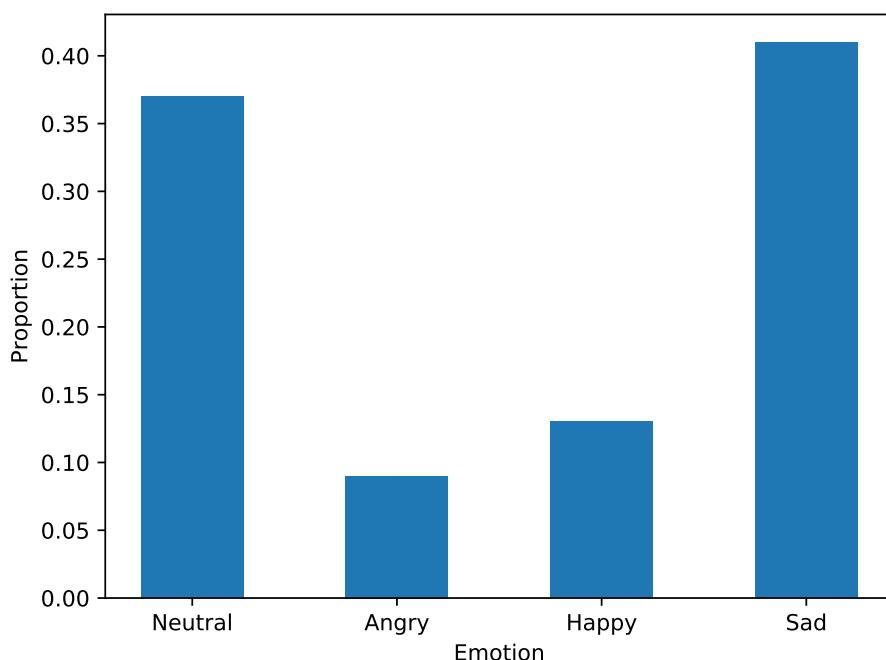


图 3.8 二分类结果中不同情感类别的分布直方图（目标情感为悲伤）

## 第4章 基于语谱图的端到端的语音情感识别

### 4.1 本章引论

上一章我们提出的基于情感对的语音情感识别框架取得了不错的效果，但是声学特征仍然采用传统的人工定义的特征，可这些特征无法保证一定能够反映语音中说话人的情感信息，所以如何从原始语音信号中抽取更为有效的特征表示仍然是一个值得关注的问题。此外，当前的情感识别都是对一个完整的句子来判断的，但声学特征通常都是对语音帧抽取的，尽管可以通过对句子中所有语音帧的声学特征计算统计函数来得到整个句子的特征表示，但这种样会丢失不同语音帧之间的时序信息。因此，如何让模型能够将这种时序信息也考虑进去同样是一个重要的问题。

近年来，深度学习的技术和工具已经飞速地发展，并且在语音信号处理领域也得到了广泛的应用。在许多的工作中，研究者逐渐开始不使用人工定义的声学特征，转而采用深度神经网络来直接从原始的语音信号上抽取相关的特征表示。这是因为深度神经网络可以从原始信号学习出更加适合任务目标的中间表示，进而使得效果提升，这被称为端到端的系统<sup>[51,52]</sup>。基于这个原因，本章将构建基于深度神经网络的端到端语音情感识别系统，为了方便神经网络处理，首先会将语音句子切分成更短的等长语音段，并抽取这些语音段的语谱图，然后通过卷积神经网络 (Convolution Neutral Network, CNN) 来从语谱图 (Spectrogram) 中抽取和情感相关的声学特征，接下来采用循环神经网络 (Recurrent Neural Network, RNN) 来建模时间序列信息，最终通过全连接神经网络建立输出和情感类别后验概率之间的映射关系，整个系统的流程如图4.1。

本章剩余的部分是这样安排的：首先我们将介绍 CNN 的结构，以及如何采用 CNN 从语谱图上抽取情感相关的特征表示，然后介绍 RNN 的结构，以及如何通过 RNN 对时间序列进行建模，最后我们将通过实验对比采用人工定义特征的方法和采用 CNN 从语谱图抽取的特征的方法在识别效果上的差异。

### 4.2 基于语谱图的卷积神经网络特征抽取

#### 4.2.1 语谱图的定义

语谱图是语音信号的不同频率的能量在时间上的变化，为了方便观测，通常会以图像的方式展示出来。下面的图4.2给出了一张语谱图，其中横轴代表时间，纵

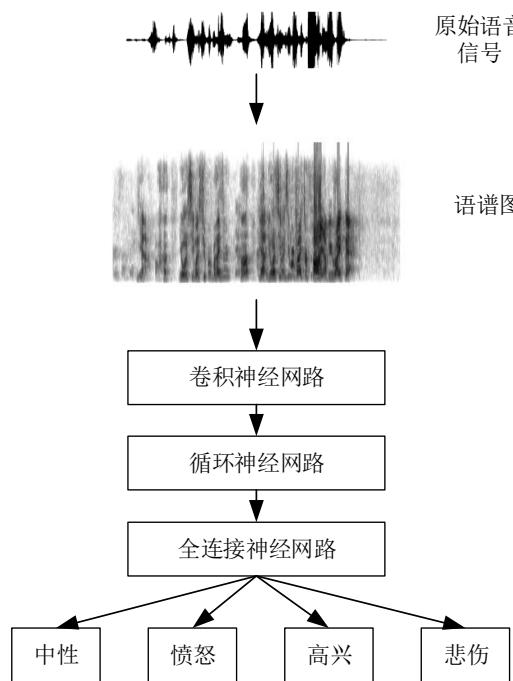


图 4.1 基于语谱图的端到端语音情感能识别系统流程图

轴代表频率，像素点的灰度值代表能量强度，颜色越深代表能量越高。假设原始语音信号为向量  $\mathbf{x}$ ，滑动窗口的长度为  $w$ ，语谱图可以通过下面的公式计算出来：

$$\text{Spectrogram}(\mathbf{x}, w) = |\text{STFT}(\mathbf{x}, w)|^2 \quad (4-1)$$

其中  $\text{STFT}(\mathbf{x}, w)$  代表对滑动窗口内的信号序列进行短时傅里叶变换 (Short-Time Fourier Transform, STFT)，前后两个窗口之间可以有重叠，每一个窗口经过傅里叶变换后得到的向量即为语谱图中每一个时间点对应的向量。当抽样频率大于语音信号频率的两倍时，根据奈奎斯特定理经过傅里叶变换的信号是能够恢复原始信号的，所以语谱图相对于原始语音信号没有任何信息损失，并且更容易观测频谱的变化。

语谱图在语音信号处理领域的运用十分广泛，通常被用来观测语音信号在频谱上的变化。在语谱图上，我们可以观测到共振峰，信号强度等信息，相比于原始的语音信号，语谱图可以更直观的观测语音在时域和频域上的变化。在本章所构建的端到端的语音情感能识别识别系统中，语谱图将作为模型的输入。

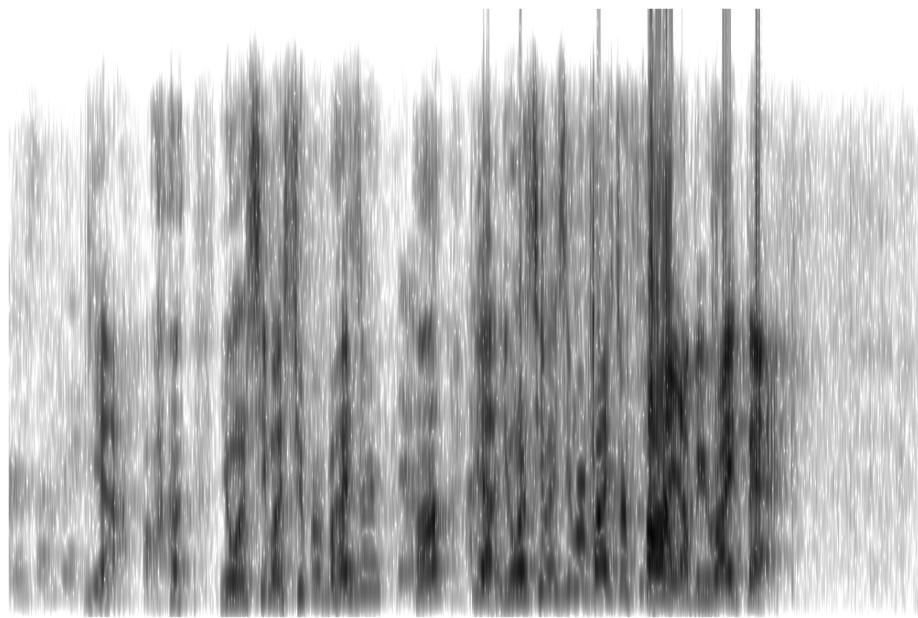


图 4.2 语谱图示例

#### 4.2.2 卷积神经网路

卷积神经网络 (Convolution Neutral Network, CNN) 是近年来被广泛使用的一种深度神经网络结构，它实际上可以被认为是将一种神经网络结构进行多份复制，然后将这些复制作用在输入的不同部分。这种网络结构可以处理规模很大的输入，同时又可以保证模型参数的数量保持不变，这使得存储模型所需的空间大大减少。这种重复利用的思想和程序设计中函数的用法很相似，就是编写一个函数的代码，然后在不同的地方调用。

为了方便表述，我们先来看一维的语音信号输入，假设我们的目的是将语音信号进行分类，输入被表示为向量  $\mathbf{x} = \{x_0, x_1, \dots, x_8\}$ ，CNN 的单一结构，也被称为卷积核，被表示为  $A$ ，每一个卷积核仅仅作用在一部分输入上，最后的分类输出通过一个全连接层  $F$  来完成，整个结构如图4.3所示。

通常一个 CNN 网络会有多个卷积核，每一个卷积核代表对一小段输入的特征表示。此外，CNN 可以堆叠多层，假设我们有一组新的卷积核  $B$ ，则可以在  $A$  上面再叠加一层  $B$ ，以此来得到更高级，更抽象的特征，结构如图4.4所示。

CNN 通常还会在后面加一层池化层 (Pooling Layer)，目的是通过计算输入特征的统计量来保证不变性，这种不变性包括平移，旋转，尺度的不变性等。此外，还可以对输入进行降维，并且使得网络能够作用在更大的输入段上。最大值池化层

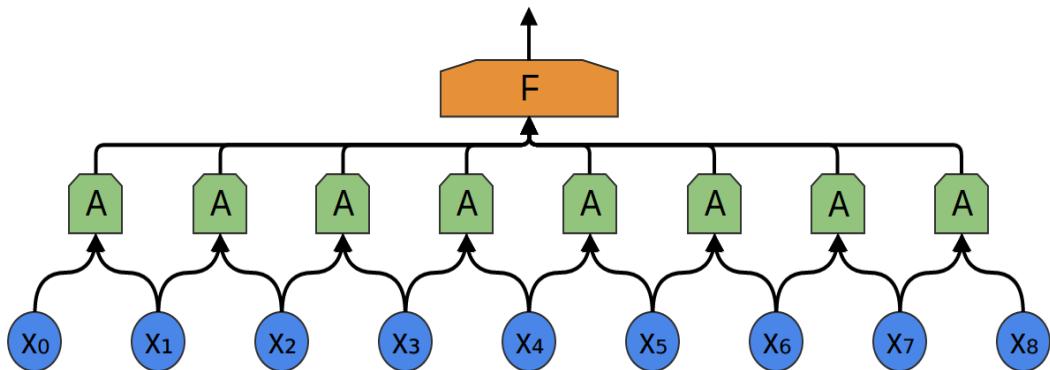


图 4.3 一维卷积神经网络（单卷积层）

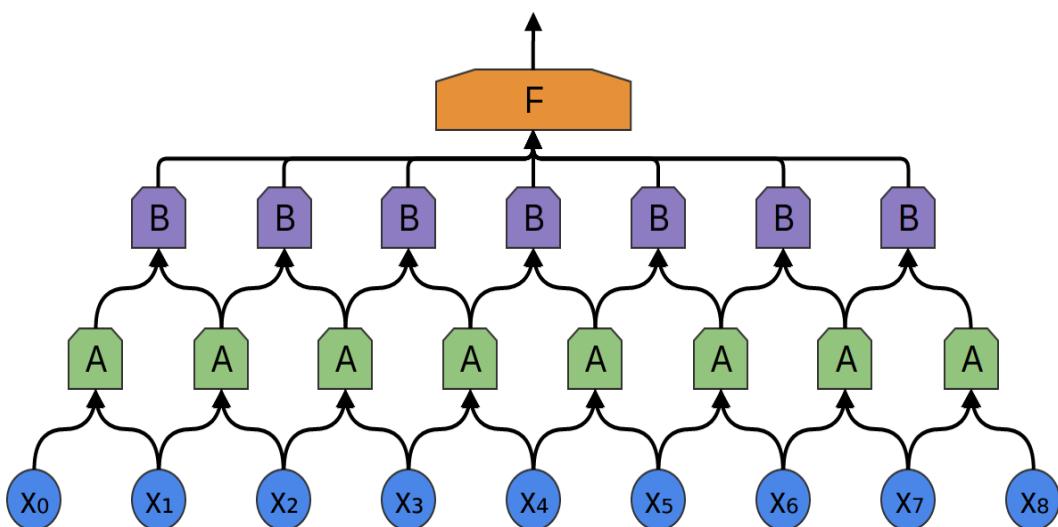


图 4.4 一维卷积神经网络（双卷积层）

(Max-Pooling Layer) 是一种被广泛使用的池化层，它的目的是从上一层的每一块输出中的找出最大值，因为通常最大值最能够反应一块输出的特性，下面的图4.5给出了最大值池化层的使用方式。

前面的例子主要介绍的是 CNN 如何处理一维的输入向量，但语谱图属于二维的输入矩阵，所以我们需要将网络结构扩展到可以处理二维输入矩阵。假设我们有输入矩阵  $\mathbf{X} = \{x_{i,j}, i = 0, 1, \dots, 8; j = 0, 1, \dots, 8\}$ ，则 CNN 的处理将会变成下面图4.6的方式。

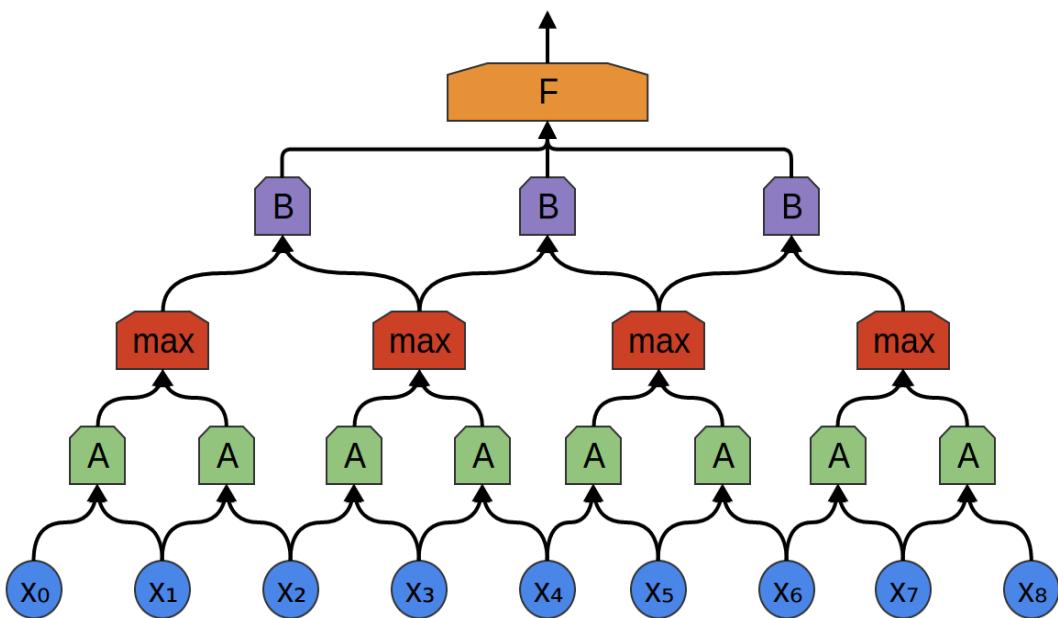


图 4.5 一维卷积神经网络（卷积层 + 池化层）

#### 4.2.3 基于语谱图的卷积神经网络特征抽取

CNN 最早用在计算机视觉领域，因为它的处理机制和人眼感知图像的机制很类似。语谱图也可以被看做是一张图像，将 CNN 作为整个深度神经网络中最开始的结构可以起到特征提取的作用。相比于传统的人工定义的声学特征，CNN 的训练是取决于最后的分类目标的，所以可以得到更加符合当前任务的特征表示。此外，由于 CNN 是整个深度神经网络模型的一部分，所以当模型中还存在 RNN 时，抽取特征时还将会考虑到时序信息。

### 4.3 基于循环神经网络的时间序列建模

#### 4.3.1 循环神经网络

传统的全连接神经网络中，所有的输入都是相互独立的，但是一些任务需要建立序列输入的前后元素之间的关系，例如在语言模型中，需要根据前面出现的词对后面的词进行预测。循环神经网络 (Recurrent Neural Network, RNN) 的目的是对序列信息进行建模，它对于序列中的每一个元素都进行相同的计算，当前的输出不仅跟当前的输入相关，而且和之前的输入也相关，下面图4.7是一个 RNN 的示意图。

图中  $\mathbf{x} = \{x_t, t = 0, 1, \dots, N\}$  为输入序列， $\mathbf{s} = \{s_t, t = 0, 1, \dots, N\}$  为隐状态序列， $\mathbf{o} = \{o_t, t = 0, 1, \dots, N\}$  为输出序列， $(U, W, V)$  是 RNN 的权重矩阵。序列中元素之

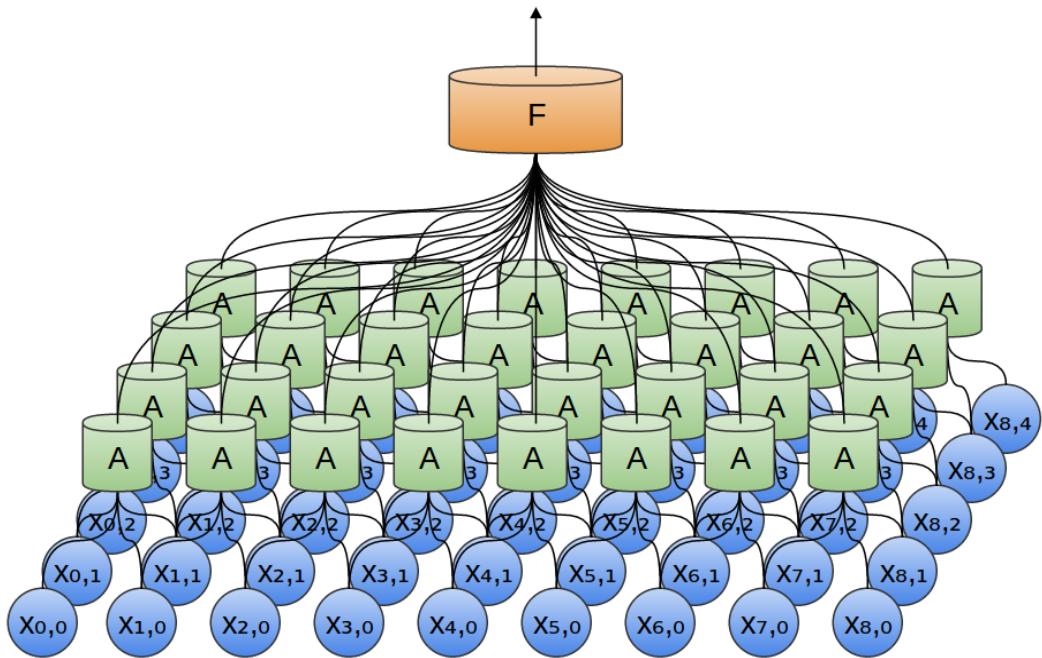


图 4.6 二维卷积神经网络（卷积层 + 池化层）

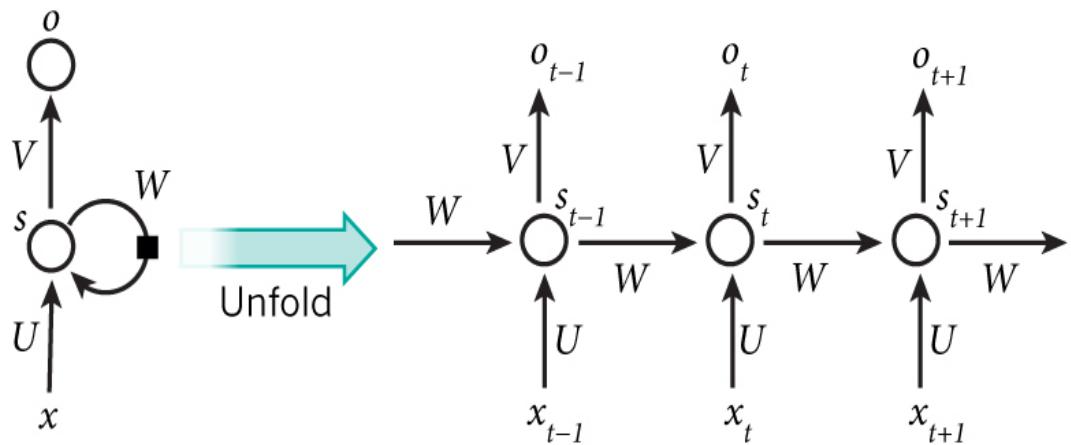


图 4.7 循环神经网络

间的映射关系可以通过下面的公式表示：

$$s_t = f(Ux_t + Ws_{t-1}) \quad (4-2)$$

$$o_t = softmax(Vs_t) \quad (4-3)$$

RNN 中的隐藏状态是能够建模序列中前后元素关系的关键，因为当前时刻的隐藏状态  $s_t$  是和上一时刻的隐藏状态  $s_{t-1}$  相关的，而当前时刻的输出  $o_t$  和当前时刻隐藏状态  $s_t$  相关的，所以这使得当前时刻的输出  $o_t$  会考虑到之前的输入信息。

对于普通的 RNN，当序列时间长度过长时，由于链式乘法规则，会导致出现梯度消失问题，从而使得网络并不能学习到太长时间的信息。这使得一些新的 RNN 网络结构被提出，例如长短时记忆 (Long-Short Time Memory) 网络，门循环单元 (Gated Recurrent Unit) 网络。它们的共同点都是引入了一种门函数 (Gated Function) 结构，使得网络在建模时序信息时，能够选择记住之前的哪些信息，这样的结构保证网络可以学习到更长时间的序列信息。此外，由于 RNN 通常只能考虑到从前到后的序列信息，但在许多任务中从后到前的序列信息也是很重要的。因此，一种双向的 RNN 结构也被提出，具体实现就是正常顺序的序列训练一个 RNN，然后将序列逆序后再训练一个 RNN，最后将两个 RNN 的输出拼接在一起，这种 RNN 结构被称为双向循环神经网络 (Bi-Directional Recurrent Neural Network)。

### 4.3.2 基于循环神经网络的时间序列建模

之前我们通过 CNN 在语谱图上抽取出了相关的特征表示，这种特征表示也是一个时间序列，每一个时间点的输入向量是原始语音信号的频谱经过抽象后得到的特征表示，所以我们可以通过 RNN 来建立不同时间步之间的时序关系。由于语音情感识别的目标是对每一个语音段判别情感类别，所以并不需要 RNN 中所有时间点的输出，仅仅只需要最后一个时间点的输出。当得到 RNN 的输出后，关于时间序列的信息已经被编码到了当前的表示中，然后再采用全连接神经网络将输出映射到每种情感类别的后验概率，我们就搭建起了整个基于深度神经网络的端到端的语音情感识别系统。

## 4.4 实验结果及分析

前面几节详细的介绍了基于深度神经网络的端到端的语音情感识别系统，下面我们将通过实验来对比采用人工定义的声学特征和在语谱图上直接抽取特征表示之间的效果差别。

### 4.4.1 实验设置

#### 4.4.1.1 情感语音数据库

本章所使用的数据库仍然为 IEMOCAP 情感语音数据库<sup>[53]</sup>，并且也只判别其中的四种情感：中性，愤怒，高兴和悲伤。与上一章不同的是我们将只考虑在对话

中诱发的情感语音，不考虑基于特定文本表演的情感语音。因为基于特定文本表演的情感语音包含太强的语言学相关的信息，并不适合分析语音中包含的情感信息。在这章的实验中我们同样会采用交叉验证 (5-Fold Cross Validation)，但是将数据库按不同的主题分为 5 个部分，每次都将 4 个部分作为训练集，另一个部分中一个说话人的数据作为验证集，其他作为测试集，下面的表格4.1是本实验中不同情感的句子数量：

表 4.1 IEMOCAP 数据库中不同情感的句子数量

中性	愤怒	高兴	悲伤	总计
1099	289	284	608	2280

#### 4.4.1.2 语谱图抽取

IEMOCAP 数据库中的语音采用 16kHz 的采样频率，每个句子的时间长度为 1s 到 30s 不等。为了方便神经网络处理，我们首先将这些句子切分成 3s 长的语音段，不足 3s 的语音段暂时保留，然后对这些语音段抽取语谱图，下面的表格4.2是抽取时的各种配置。

表 4.2 语谱图抽取的参数配置

参数名称	参数值
滑动窗口类型	汉明窗 (Hamming Window)
滑动窗口长度	40ms
滑动窗口偏移	10ms
频谱范围	0-4kHz
频谱分辨率	800

在抽取完语谱图后，我们将通过取对数降低能量表示的取值范围，然后采用 z-normalization 对所有的样本进行正则化，其中均值为语谱图中每一个频率刻度在所有样本中的所有时间的能量的均值，标准差的计算也是如此。在完成正则化后，还需要将不足 3s 的语谱图通过在末尾加 0 补齐到 3s，这样就可以保证所有输入样本的长度是相同的。我们最后输入神经网络模型的将是一个  $300 \times 400$  的矩阵。

#### 4.4.1.3 深度神经网络结构和参数

在实验中我们采用 CNN 从语谱图中抽取特征表示，然后使用 RNN 建模语音信号的时序信息，最后通过全连接网络映射到情感类别的后验概率。在尝试了多种的神经网络结构和参数后，下面的图4.8是我们当前最好效果的网络结构。此外，由于不同情感的句子数量有所差异，所以在训练时需要对误差函数 (Loss Function) 乘以权重，这个权重和样本对应的情感类别在数据库中的句子数量成反比，这样可以保证模型不会对数量样本多的情感产生偏好。还有就是我们最后需要得到整个句子的情感类别，所以将会对句子切分的所有语音段计算不同情感的后验概率，然后将所有语音段的输出结果取平均值，最后将平均值最大的那种情感作为最后的识别结果。

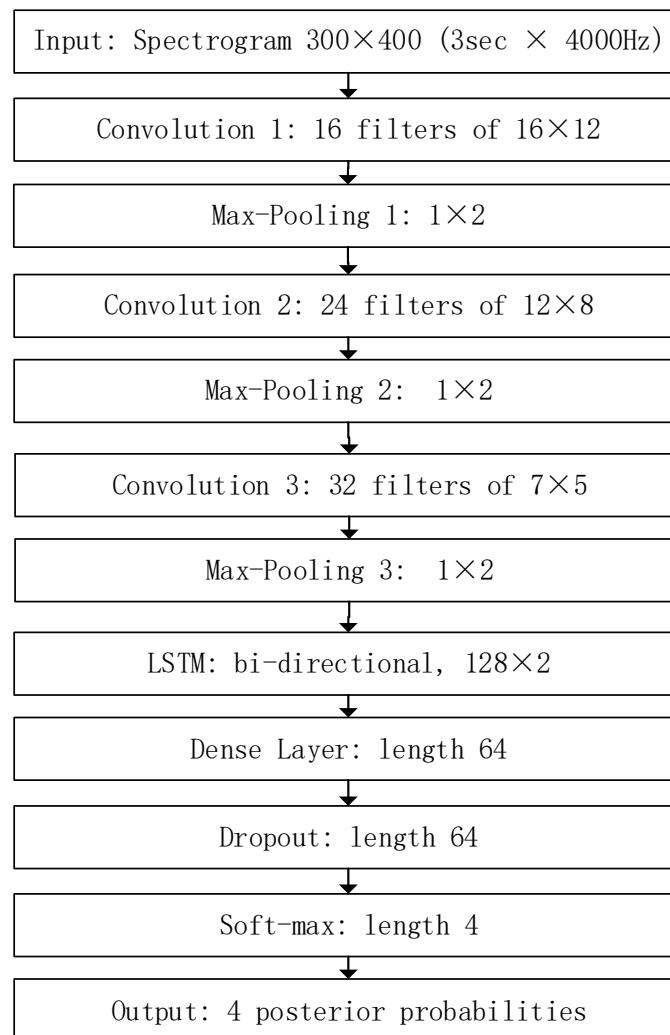


图 4.8 深度神经网络结构

#### 4.4.2 实验结果

实验结果主要可以分为两个部分，第一部分主要比较了基于人工定义特征的方法和基于语谱图抽取特征的方法之间识别率的差异；第二部分主要通过图像来展示语谱图通过 CNN 后得到的特征表示，从而进一步证明这种抽取特征的方式能够关注到语谱图中的一些特定模式信息。

##### 4.4.2.1 准确率对比

本次实验的评价指标仍然是加权准确率 (Weighted Accuracy, WA) 和不加权准确率 (Unweighted Accuracy, UA)，在下面的表4.3中，我们展示了以前采用传统声学特征的工作中在 IEMOCAP 数据集上取得的最好的结果，同样的特征集在我们设计的神经网络模型上的结果，以及语谱图作为输入在我们神经网络模型上的结果。这里采用的声学特征集合和上一章相同，也是 Interspeech2009 语音情感识别比赛的特征集合，共有 384 维特征。其中，“最佳结果”代表之前在传统声学特征集上取得的最好的结果，“神经网络”代表在本章设计的神经网络模型下的结果。从实验结果中我们可以看出，当通过深度神经网络直接在语谱图上抽取特征表示时，相比于采用传统声学特征的方法，不仅在 WA 上取得了明显的提升，而且在 UA 上也取得了相近的结果。这证明采用神经网络在语谱图上抽取的特征表示相比于传统声学特征更能够反映语音中的情感信息。

表 4.3 不同方法的准确率

模型	WA	UA
传统声学特征（最佳结果）	63.90%	62.80%
传统声学特征（神经网络）	62.41%	60.02%
语谱图（神经网络）	<b>67.30%</b>	<b>62.00%</b>

##### 4.4.2.2 基于卷积神经网络的特征表示

为了进一步观察深度神经网络在语谱图上学习到了什么，我们将最后一层 CNN 的输出以图像的方式展示出来。下面图4.9和图4.10展示了一段中性语音的语谱图和最后一层 CNN 中两个卷积核的输出值，其中横轴代表时间，纵轴代表频率，红色代表高能量，蓝色代表低能量。从图4.10中 (a) 的激活输出中可以观察到这个卷积核学习到了语谱图中垂直方向的模式信息，而从 (b) 的激活输出中则可

以观察到另一个卷积核学习到了语谱图水平方向的模式信息。这些实验数据表明 CNN 可以从原始语谱图中捕捉到关于语音时域和频域的相关信息。

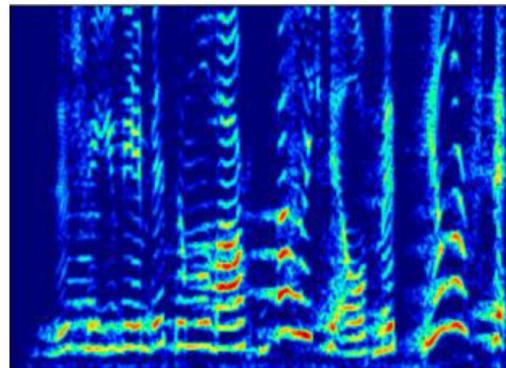
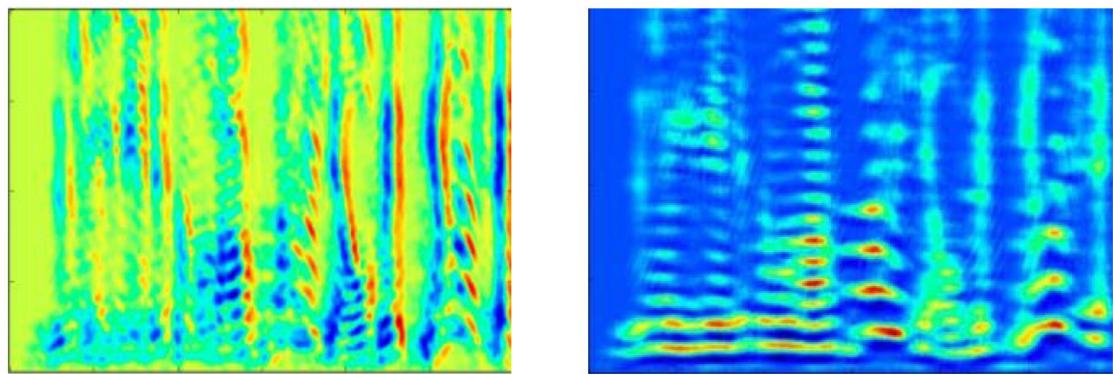


图 4.9 原始语谱图



(a) 垂直方向模式

(b) 水平方向模式

图 4.10 卷积神经网络激活输出

## 4.5 本章小结

本章我们介绍了基于深度神经网络的端到端的语音情感识别系统，首先通过 CNN 来直接从语谱图上抽取和情感相关的特征表示，然后采用 RNN 来建模语音信号的时序关系，最后通过一个全连接神经网络将 RNN 最后一个时间步的输出映射到不同情感的后验概率。在情感语音数据库 IEMOCAP 上，相比于之前基于传统声学特征的方法，基于语谱图的端到端系统可以取得更好的识别率。

## 第5章 基于变长语音段的情感识别

### 5.1 本章引论

上一章我们介绍了基于深度神经网络的端到端的语音情感识别系统，但是为了方便神经网络模型处理，一个完整的语音句子被切分成了更小的等长语音段，每一个语音段都被标记为对应句子的情感类别。这样的处理方式会引入一些问题，因为在我们实际的观测中，一个非中性的句子中并不是所有的部分都会包含有明显的情感信息，往往只有一部分语音包含情感信息。当我们把句子切分成多个语音段时，可能并不是所有的语音段都包含有情感信息，但仍然会将不包含情感信息的语音段标记为句子的情感类别。这样的数据在训练模型时会对模型造成混淆，因为同样的中性语音段，有的被标记为中性，有的被标记为其他情感，使得模型无法区分哪些应该是真正的中性语音。除了对中性情感的识别会产生影响以外，由于在预测时是句子划分的所有语音段共同来决定句子最终的情感类别，而带情感的句子的中性语音段之间区别也不大，所以也有可能会导致在识别时不同情感之间发生混淆。

为了进一步验证我们对于句子中只有部分的语音段包含有明显的情感信息这一猜测，下面图5.1展示了一个标记为愤怒的句子中前后两个语音段的语谱图，其中横轴代表时间，纵轴代表频率，颜色越深代表能量越高，(a) 属于句子前半部分的，(b) 属于句子后半部分。通常来说，愤怒的语音在各个频段的能量都相对较高，可以看到图中(b) 的语音段的能量较高，看上去更像一段愤怒的语音，而(a) 的语音段则看上去却不像愤怒语音。通过人工听测这两个语音段，同样也发现(b) 的语音段听上去包含有更多的愤怒情感，(a) 的语音段听上去却更像中性语音。然而，当我们直接听整个句子时，却发现前面的中性语音段可以增强后面的愤怒语音段的愤怒感觉，也就是所谓的欲扬先抑。但是句子的长度通常都是不相同的，所以我们认为设计一种可以处理变长语音段的神经网络结构，并将整个句子作为输入能够更好地提升模型的识别率。

本章剩余的部分是这样安排的：首先我们将介绍变长语音段如何抽取语谱图，然后将介绍如何实现能够处理变长输入的深度神经网络模型，最后将通过实验比较定长输入的神经网络和变长输入的神经网络在识别效果上的差异。

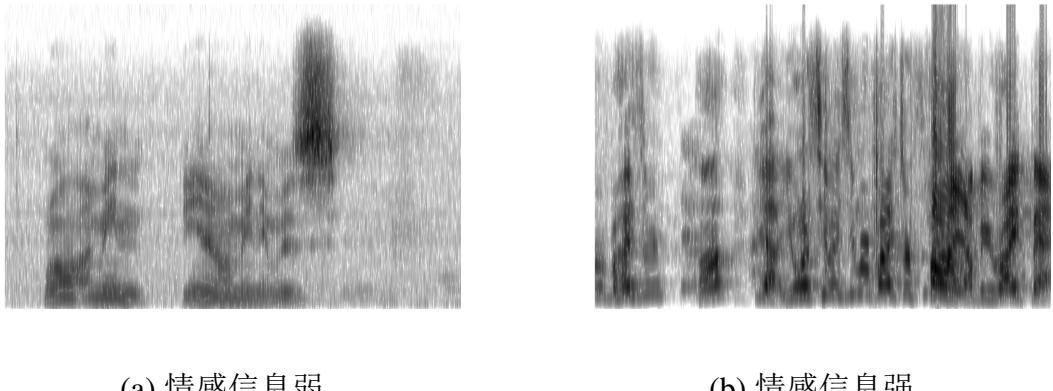


图 5.1 愤怒句子不同语音段的语谱图

## 5.2 变长语音段的语谱图抽取

本章采取的语谱图抽取参数和上一章相同，唯一不同的是我们将对整个句子进行语谱图抽取，而不是对切分的等长语音段。由于在进行神经网络训练时，通常会将多个样本组成一个批次 (Batch)，然后一起放进模型调整参数，同时，一个批次中的数据长度需要相同，但不同长度的句子得到的语谱图的长度各不相同，所以需要通过将一个批次中的语谱图都用 0 补齐到最长语谱图的长度。为了提升计算效率，所有的语谱图将会按照时间长度进行排序，然后将时间长度接近的语谱图放入一个批次，这样可以保证需要补齐的 0 最少，从而可以节约存储空间，增加计算速度。

## 5.3 变长神经网络结构

在上一章基于深度神经网络的端到端的语音情感识别系统中，我们主要采用了两种神经网络结构：卷积神经网络 (Convolution Neutral Network, CNN) 和循环神经网络 (Recurrent Neural Network, RNN)。这两种神经网络结构通常的用法都是处理定长的输入，但它们同样也具备处理变长输入的能力。为了方便描述，假设输入序列为  $\mathbf{s} = \{x_1, x_2, \dots, x_V, \dots, x_T\}$ ，其中  $\mathbf{s}_1 = \{x_1, x_2, \dots, x_V\}$  是有效的部分， $\mathbf{s}_2 = \{x_{V+1}, x_{V+2}, \dots, x_T\}$  是补齐的 0。我们设计神经网络的主要目的就是不让补齐的 0 对最终的结果产生影响，下面将介绍本文是如何实现变长的 CNN 和 RNN 的。

### 5.3.1 变长卷积神经网络

卷积神经网络 (Convolution Neutral Network, CNN) 是通常是处理定长的输入，特别是在计算机视觉领域，但 CNN 本身只是在训练卷积核，而这些卷积核和输入

的大小是无关的，因此 CNN 具备处理变长输入的能力。为了只保留有效的输入  $s_1$  对输出结果的影响，我们将采用掩码 (Masking) 去屏蔽补齐的无效输入  $s_2$  所产生的输出，如下面的公式所示：

$$\mathbf{s}_{Conv} = Conv(\mathbf{s}) \bullet Mask(\mathbf{s}) \quad (5-1)$$

其中  $Conv(\mathbf{s})$  代表序列  $s$  在经过卷积层后的输出， $Mask(\mathbf{s})$  代表掩码矩阵，矩阵中代表有效部分的位置被置为 1，无效部分的位置被置为 0，最终的有效输出为  $\mathbf{s}_{Conv} = \{y_1, y_2, \dots, y_V, \dots, y_T\}$ ，是  $Conv(\mathbf{s})$  和  $Mask(\mathbf{s})$  通过点乘运算 (Element-Wise Multiply) 得到的。此外，卷积层后面的通常都会连接一个池化层 (Pooling Layer)，我们还需要关注有效部分和无效部分的边界，因为这里有可能会引入无效的信息。例如，假设  $\mathbf{s}_{Conv}$  是最大值池化层 (Max-Pooling Layer) 的输入，如果池化层的单元大小为 2，一个单元的输入为  $y_V$  和  $y_{V+1}$ ，则当  $y_V < 0$  且  $y_{V+1} = 0$  时，这个单元的输出为 0，然而我们期望的输出应该是  $y_V$ ，因为  $y_{V+1}$  是无效部分得到的输出。在实际的测试中，我们发现这样无效信息的引入会导致神经网络无法收敛。因此， $y_V$  所在位置的掩码也会被设置为 0，这样输入最大值池化层时就不会引入无效的信息了。通过加入掩码矩阵的方法，我们可以保证补齐的 0 不会对 CNN 的输出产生影响，这可以保证相同输入在训练和测试阶段可以得到相同的输出，因为在测试阶段样本是不会用 0 补齐的。

### 5.3.2 变长循环神经网络

循环神经网络 (Recurrent Neural Network, RNN) 是用来建模语音信号的时序信息的，对于每一个时间步的输出都采用相同的参数进行计算，同样也可以用来处理变长的序列输入。由于语音情感识别输入序列分类任务，所以我们只需要 RNN 最后一个时间步的输出。假设  $\mathbf{s}$  是 RNN 的输入，期望的结果是时间步  $t = V$  时的输出，所以我们可以忽略  $t = V + 1$  以后的所有输出，保证无效的补齐部分不会对输出产生影响。此外，对于双向的 RNN，反向的 RNN 将会输出  $t = 0$  时的输出，所以最后输出会将正向和反向的结果拼接在一起。

## 5.4 实验结果及分析

前面几节详细的介绍了如何设计深度神经网络来处理变长的语音段，下面我们将通过实验来对比采用上一章的定长方法和本章的变长方法之间的效果差别。

### 5.4.1 实验设置

数据库同样是采用 IEMOCAP 情感语音数据库<sup>[53]</sup>, 其他实验设置和上一章的相同, 唯一不同的就是模型的语谱图输入在时域上不是固定大小了, 而是和语音句子的长度相关。下面图 2 是通过实验得到的最佳神经网络结构, 其中  $T$  是语音句子的时间长度,  $N$  是语谱图的时域长度。

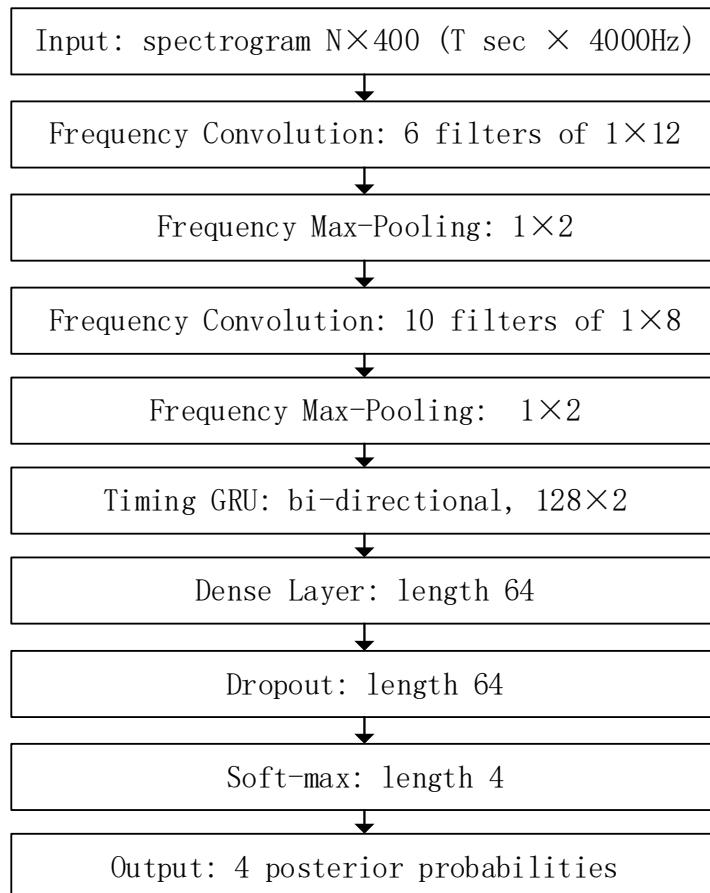


图 5.2 深度神经网络结构

### 5.4.2 实验结果

实验结果主要可以分为两个部分, 第一部分主要比较了定长方法和本章的变长方法之间识别率的差异; 第二部分主要通过图像来展示 RNN 不同节点在不同时间步的激活程度, 从而进一步证明变长的方法可以减轻模型对于不同情感的混淆。

#### 5.4.2.1 准确率对比

本次实验的评价指标仍然是加权准确率 (Weighted Accuracy, WA) 和不加权准确率 (Unweighted Accuracy, UA), 在下面的表5.1中, 我们展示了上一章定长方法

取得的最好的结果，由于本章采用的网络结构和上一章不同，所以也会展示定长的输入在本章神经网络模型上的结果，以及变长的输入在本章神经网络上的结果，其中，“最佳结果”代表上一章的最好的结果，“定长模型”和“变长模型”分别代表在本章设计的神经网络模型下定长输入和变长输入得到的结果。从实验结果中我们可以看出，当采用变长的神经网络结构时，相比于定长的方法，在 WA 和 UA 上均得到了提升。这证明采用变长的输入比切分成定长的输入能够取得更好的效果。

表 5.1 不同方法的准确率

模型	WA	UA
最佳结果	67.30%	62.00%
定长模型	68.86%	57.45%
变长模型	<b>71.45%</b>	<b>64.22%</b>

为了进一步分析实验结果，我们在下面的表格5.2和表格5.3中分别展示了定长输入和变长输入在本章的神经网络结构中得到的混淆矩阵。我们可以看到中性情感在变长方法的识别率提升了，正如我们在5.1节分析的，当整个句子输入模型时可以减轻中性情感和其他情感的混淆。此外，高兴的准确率也得到了显著的提升，这或许是因为定长方法中，高兴句子切分出的语音段大多为中性情感，所以会被错分为其他情感。还有就是悲伤的识别率降低了，这或许是因为其他情感的准确率提升了，所以导致一些悲伤的语音被错分到其他情感。这些结果可以证明变长的方法确实能够提升部分情感的识别率。

表 5.2 混淆矩阵（定长输入）

实际 \ 预测	中性	愤怒	高兴	悲伤
中性	<b>71.75%</b>	8.88%	5.93%	13.45%
愤怒	30.84%	<b>58.79%</b>	7.05%	3.34%
高兴	55.92%	31.42%	<b>11.72%</b>	0.95%
悲伤	11.41%	0%	1.02%	<b>87.57%</b>

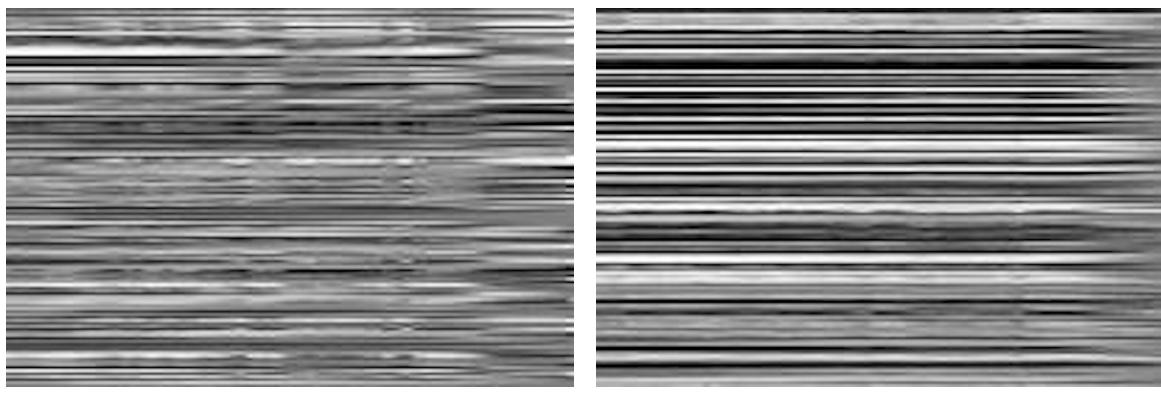
#### 5.4.2.2 循环神经网络输出

通过观察神经网络的激活输出能够验证网络学习到的信息，下面的图5.3展示了一个中性语音的句子通过神经网络后 RNN 不同节点的激活输出。左边的图片代

表 5.3 混淆矩阵（变长输入）

实际 \ 预测	中性	愤怒	高兴	悲伤
中性	<b>73.64%</b>	2.74%	12.41%	11.21%
愤怒	11.44%	<b>59.55%</b>	26.52%	2.5%
高兴	45.20%	13.81%	<b>40.05%</b>	0.95%
悲伤	15.89%	0%	0.48%	<b>83.64%</b>

表定长神经网络，右边的图片代表变长神经网络，其中横轴代表时间，纵轴代表 RNN 不同的节点，颜色越深代表激活程度越高。从图中可以观察到右边图片中的条纹比左边图片中的要清晰很多。此外，我们也发现特定的情感在特定的节点上有更高的激活值。图5.4展示了三种不同情感的语音在变长网络中的 RNN 的不同节点的激活输出，红色方框中的是高激活度的节点。从图中可以发现高兴和愤怒的激活节点比较相似，因为这两种情感的声学表现也比较相似，而悲伤地激活节点和高兴的激活节点就有明显的不同。不过中性情感无法找到特定的激活节点，这可能是因为中性语音没有一种特别明显的声学表现。这些不同情感对应的激活节点可以在变长网络中清楚地观察到，但在定长网络中，由于条纹非常模糊，所以很难观察到这样的现象。这些实验结果进一步证明了变长的方法能够减轻模型对不同情感的混淆。



(a) 定长方法

(b) 变长方法

图 5.3 定长方法和变长方法中循环神经网络的激活输出

## 5.5 本章小结

本章主要设计了一种能够处理变长语音段输入的深度神经网络结构，相比于上一章提出的定长输入的深度神经网络模型，可以有效地减轻模型对不同情感的

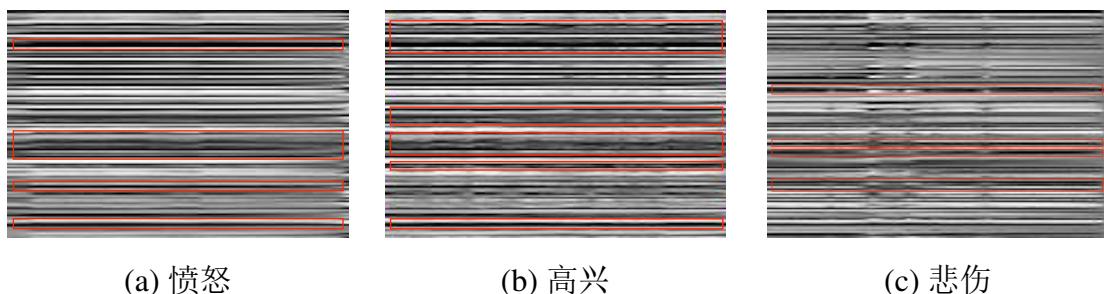


图 5.4 不同情感在变长循环神经网络上的激活节点

混淆。在情感语音数据库 IEMOCAP 上，变长的方法方法相比于定长的方法能够取得更好的识别率。

## 第6章 总结与展望

### 6.1 本文工作总结

随着人机交互日应用的日益普遍，语音情感能识别正在逐渐受到更多研究者的重视。本文主要对语音情感能识别中特征抽取和选择的方式进行了研究，包括基于传统声学特征的语音情感能识别框架和基于深度学习的端到端的语音情感能识别框架。针对现有研究中存在的关于特征选取，特征抽取和变长语音段输入相关的问题，提出了一些有效的解决方案。本文主要的工作可以分为以下几个部分：

**一、提出一种基于情感对的语音情感能识别框架，为不同的情感对选择不同的声学特征，并在最后的决策融合过程中引入心理学的情感空间模型，从而提升了系统的识别率。**传统的语音情感能识别系统通常为所有的情感选取相同的声学特征来完成最后的情感分类，但实验证明不同的情感和不同的声学特征的相关性并不同。针对这一问题，我们将分别为不同的情感对选取不同的特征集合，将原先的多分类问题转变为多个二分类问题，并在最后的决策融合过程中通过贝叶斯分类器引入情感空间的信息。在公开的情感语音数据集 IEMOCAP 上，我们方法取得了比传统的识别框架更好的准确率。

**二、构建了基于深度神经网络的端到端的语音情感能识别系统，使用语谱图代替传统的声学特征，从而提升了系统的识别率。**随着深度学习技术和工具的发展，许多的研究者开始采用深度神经网络在原始语音信号上直接构建分类或者回归模型，被称之为端到端的系统。相比于采用传统的声学特征，这种方法可以抽取到更符合任务目标的特征表示。语谱图是语音信号的一种无损表示，我们通过卷积神经网络来从语谱图上直接抽取和情感相关的特征表示，然后通过循环神经网络来建模语音信号的时序信息，最后通过全连接网络将输出映射到不同情感的后验概率。相比于采用传统的声学特征，端到端的系统能够取得更好的准确率。

**三、设计了一种能够处理变长语音段的神经网络结构来实现端到端的系统，消除了语音分段带来的中性语音和情感语音的混淆，从而提升了系统的识别率。**在使用深度神经网络实现端到端的语音情感能识别系统时，由于卷积神经网络和循环神经网络很难处理变长的输入，通常会把变长的语音句子切分成多段等长的语音段，然后将所有语音段都标记为对应句子的情感标签，但这样会导致部分中性语音段被标记为有情感。针对这一问题，我们采用补齐和掩码的方式来处理神经网络中变长的输入序列，避免了错误标注带来了的效果变差。相对于切分等长语音段的方法，我们直接输入整个变长语音句子的方法可以在相同的数据集上取得更

好的准确率。

## 6.2 未来工作展望

我们的工作已经探索一些关于情感相关的特征抽取和选择的问题，包括传统的机器学习模型和深度学习模型，但在现有工作的基础上仍然有许多问题可以继续进行研究。

在提出的基于情感对的语音情感识别框架中，我们只采用比较小的声学特征集合来进行特征选择，有很多的声学特征都没有考虑进来。如果将声学特征集合扩大，本文采用的特征选择算法会需要较长的时间来执行，所以可以进一步采用更为快速的特征选择算法，例如遗传算法等。此外，本文只采用一些比较简单的分类模型来预测情感，但仍然有许多其他的分类模型没有被测试过，因此系统的识别率仍然有很大的提升空间。

在我们提出的基于深度学习的端到端语音情感识别系统中，也只是使用了 CNN、RNN 等几种深度神经网络结构。现在已经有许多新的神经网络结构被提出，并且在其他的一些序列分类问题上取得了很不错的成绩，例如注意力的机制，这些新的神经网络结构或许可以进一步提升系统的识别率。此外，情感语音数据的获取比较困难，所以训练数据比较少，而深度学习模型通常都有比较多的参数，需要大量的训练数据来学习参数，只使用少量数据会导致模型训练时过拟合，所以我们需要设计一些无监督的方法来增加训练数据，例如先用少量的标记数据去训练一个简单模型，然后去筛选未标记的数据。除了增加数据以外，也可以通过经验加入一些规则，从而能够较少模型的学习负担。

## 参考文献

- [1] Ayadi M E, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases[M]. [S.l.]: Elsevier Science Inc., 2011: 572–587
- [2] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1): 37–50.
- [3] Cowie R, Douglasscowie E, Tsapatsoulis N, et al. Emotion recognition in human-computer interaction[J]. IEEE Signal Processing Magazine, 2002, 18(1): 32–80.
- [4] Lee C M, Narayanan S S. Toward detecting emotions in spoken dialogs[J]. IEEE Transactions on Speech & Audio Processing, 2005, 13(2): 293–303.
- [5] Murray I R, Arnott J L. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion[J]. Journal of the Acoustical Society of America, 1993, 93(2): 1097–1108.
- [6] Atal B S. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification.[J]. Journal of the Acoustical Society of America, 1974, 55(6): 1304–22.
- [7] Banse R, Scherer K R. Acoustic profiles in vocal emotion expression[J]. Journal of Personality & Social Psychology, 1996, 70(3): 614–636.
- [8] Bou-Ghazale S E, Hansen J H L. A comparative study of traditional and newly proposed features for recognition of speech under stress[J]. Speech & Audio Processing IEEE Transactions on, 2000, 8(4): 429–442.
- [9] Hernando J, Nadeu C. Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition[J]. Speech & Audio Processing IEEE Transactions on, 1997, 5(1): 80–84.
- [10] Kaiser L. Communication of affects by single vowels[J]. Synthese, 1962, 14(4): 300–319.
- [11] Rabiner L, Juang B H. Fundamentals of speech recognition[M]. [S.l.]: Tsinghua University Press, 1999: 353–356
- [12] Davitz J R. The communication of emotional meaning.[M]. [S.l.]: McGraw-Hill, 1964
- [13] Rabiner L R, Juang B H. An introduction to hidden markov models[J]. Current protocols in bioinformatics / editorial board, Andreas D. Baxevanis ... [et al.], 2007, Appendix 3(Appendix 3): Appendix 3A.
- [14] Ephraim Y, Merhav N. Hidden markov processes[J]. IEEE Transactions on Information Theory, 2002, 48(6): 1518–1569.
- [15] Schuller B, Rigoll G, Lang M. Hidden markov model-based speech emotion recognition[C]// International Conference on Multimedia and Expo, 2003. ICME '03. Proceedings. [S.l.: s.n.], 2003: I–401–4 vol.1.
- [16] Nogueiras A, Moreno A, Bonafonte A, et al. Speech emotion recognition using hidden markov models[C]//Eurospeech 2001 Scandinavia, European Conference on Speech Communication and Technology, INTERSPEECH Event, Aalborg, Denmark, September. [S.l.: s.n.], 2012: 2679–2682.
- [17] Kwon O W. Emotion recognition by speech signals[J]. EUROSPEECH-2003, 2003: 125–128.

- [18] Lee C M. Emotion recognition based on phoneme classes[J]. Proc. ICSLP, Oct. 2004, 2004: 889–892.
- [19] Vlassis N, Likas A. A greedy em algorithm for gaussian mixture learning[J]. Neural Processing Letters, 2002, 15(1): 77–87.
- [20] Reynolds D A, Rose R C. Robust text-independent speaker identification using gaussian mixture speaker models[J]. IEEE Trans Speach & Audio Processing, 1995, 3(1): 72–83.
- [21] Rissanen J. Modeling by shortest data description[J]. Automatica, 1978, 14(5): 465–471.
- [22] El-Yazeed M F A, Gamal M A E, Ayadi M M H E. On the determination of optimal model order for gmm-based text-independent speaker identification[J]. Eurasip Journal on Advances in Signal Processing, 2004, 2004(8): 1–10.
- [23] Vlassis N, Likas A. A kurtosis-based dynamic approach to gaussian mixture modeling[M]. [S.l.]: IEEE Press, 1999: 393–399
- [24] Breazeal C, Aryananda L. Recognition of affective communicative intent in robot-directed speech[J]. Autonomous Robots, 2002, 12(1): 83–104.
- [25] Slaney M, Mcroberts G. Baby ears: a recognition system for affective vocalizations[M]. [S.l.]: Elsevier Science Publishers B. V., 2003
- [26] Burges C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining & Knowledge Discovery, 2008, 2(2): 121–167.
- [27] Schuller B, Rigoll G, Lang M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. [S.l.: s.n.], 2004: I-577–80 vol.1.
- [28] Pierre-Yves O. The production and recognition of emotions in speech: features and algorithms [M]. [S.l.]: Academic Press, Inc., 2003: 423–423
- [29] Carpenter G A. Neural network models for pattern recognition and associative memory[M]. [S.l.]: Elsevier Science Ltd., 1989: 243–257
- [30] Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks [C]//International Conference on Neural Information Processing, 1999. Proceedings. ICONIP. [S.l.: s.n.], 2000: 495–501 vol.2.
- [31] Crystal T. Linear prediction of speech[J]. Proceedings of the IEEE, 1978, 66(2): 266–267.
- [32] Hozjan V, Kai Z. Context-independent multilingual emotion recognition from speech signals [J]. International Journal of Speech Technology, 2003, 6(3): 311–320.
- [33] Petrushin V A. Emotion recognition in speech signal: Experimental study, development, and application[C]//The Proceedings of the. [S.l.: s.n.], 2000: 222–225.
- [34] Schuller B, Lang M, Rigoll G. Robust acoustic speech emotion recognition by ensembles of classifiers[C]//31. Jahrestagung Für Akustik, Daga. [S.l.: s.n.], 2005: 329–330.
- [35] Kuncheva L I. Combining pattern classifiers: Methods and algorithms[M]. [S.l.]: Wiley-Interscience, 2004
- [36] Lugger M, Janoir M E, Yang B. Combining classifiers with diverse feature sets for robust speaker independent emotion recognition[C]//Signal Processing Conference, 2009 European. [S.l.: s.n.], 2015: 1225–1229.

- [37] Mashao D J, Skosan M. Combining classifier decisions for robust speaker identification[J]. *Pattern Recognition*, 2006, 39(1): 147–155.
- [38] Wu J, Mullin M D, Rehg J M. Linear asymmetric classifier for cascade detectors[C]// *International Conference on Machine Learning*. [S.l.: s.n.], 2005: 988–995.
- [39] Kuncheva L I. A theoretical study on six classifier fusion strategies[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2002, 24(2): 281–286.
- [40] Han K, Yu D, Tashev I. Speech emotion recognition using deep neural network and extreme learning machine[C]// *INTERSPEECH*. [S.l.: s.n.], 2014.
- [41] Lee J, Tashev I. High-level feature representation using recurrent neural network for speech emotion recognition[M]. [S.l.: s.n.], 2015.
- [42] Huang Z, Dong M, Mao Q, et al. Speech emotion recognition using cnn[C]// *ACM International Conference on Multimedia*. [S.l.: s.n.], 2014: 801–804.
- [43] Le D, Provost E M. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks[C]// *Automatic Speech Recognition and Understanding*. [S.l.: s.n.], 2013: 216–221.
- [44] Rana R. Emotion classification from noisy speech - a deep learning approach[M]. [S.l.: s.n.], 2016.
- [45] Chernykh V, Sterling G, Prihodko P. Emotion recognition from speech with recurrent neural networks[M]. [S.l.: s.n.], 2017.
- [46] Jaitly N, Hinton G. Learning a better representation of speech soundwaves using restricted boltzmann machines[C]// *IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2011: 5884–5887.
- [47] Bhargava M, Rose R. Architectures for deep neural network based acoustic models defined over windowed speech waveforms[C]// *INTERSPEECH*. [S.l.: s.n.], 2015.
- [48] Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]// *IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2015: 4580–4584.
- [49] Amodei D, Anubhai R, Battenberg E, et al. Deep speech 2: End-to-end speech recognition in english and mandarin[J]. *Computer Science*, 2015.
- [50] Variani E, Lei X, Mcdermott E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]// *IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2014: 4052–4056.
- [51] Trigeorgis G, Ringeval F, Brueckner R, et al. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network[C]// *IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2016.
- [52] Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms[C]// *INTERSPEECH*. [S.l.: s.n.], 2017: 1089–1093.
- [53] Busso C, Bulut M, Lee C C, et al. Iemocap: interactive emotional dyadic motion capture database[J]. *Language Resources & Evaluation*, 2008, 42(4): 335.

- [54] Van Bezooijen R, Otto S A, Heenan T A. Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics.[J]. Journal of Cross-Cultural Psychology, 1983, 14(4): 387–406.
- [55] Ortony A, Turner T J. What's basic about basic emotions?[J]. Psychological Review, 1990, 97 (3): 315–31.
- [56] Robinson M D, Watkins E R, Harmonjones E. Handbook of cognition and emotion: volume 22 [M]. [S.l.: s.n.], 2013: 237–244.
- [57] Campbell N. Databases of emotional speech[J]. Proceedings of Isca Itrw on Speech & Emotion Developing A Conceptual Framework, 2000.
- [58] Mazuka R, Igarashi Y, Martin A, et al. Emotions and speech: some acoustical correlates[J]. Laboratory Phonology, 2015, 52(1): 1238–1250.
- [59] Batliner A, Fischer K, Huber R, et al. Desperately seeking emotions or: Actors, wizards, and human beings[J]. Isca Tutorial & Research Workshop on Speech & Emotion, 2000.
- [60] Johnstone T, Van Reekum C M, Hird K, et al. Affective speech elicited with a computer game. [J]. Emotion, 2005, 5(4): 513.
- [61] Burkhardt F, Paeschke A, Rolfs M, et al. A database of german emotional speech[C]// INTERSPEECH 2005 - Eurospeech, European Conference on Speech Communication and Technology, Lisbon, Portugal, September. [S.l.: s.n.], 2005: 1517–1520.
- [62] Hansen J H L, Bou-Ghazale S E. Getting started with susas: a speech under simulated and actual stress database[C]//European Conference on Speech Communication and Technology, Eurospeech 1997, Rhodes, Greece, September. [S.l.: s.n.], 1997.
- [63] Morrison D, Wang R, Silva L C D. Ensemble methods for spoken emotion recognition in call-centres[J]. Speech Communication, 2007, 49(2): 98–112.
- [64] Liberman M, Davis K, Grossman M, et al. Emotional prosody speech and transcripts[M]. [S.l.: s.n.].
- [65] Engberg I S, Hansen A V, Andersen O, et al. Design, recording and verification of a danish emotional speech database[C]//European Conference on Speech Communication and Technology, Eurospeech 1997, Rhodes, Greece, September. [S.l.: s.n.], 1997.
- [66] Hozjan V, Kacic Z, Moreno A, et al. Interface databases: design and collection of a multilingual emotional speech database[M]. [S.l.: s.n.], 2002.
- [67] Schuller B, Reiter S, Muller R, et al. Speaker independent speech emotion recognition by ensemble classification[C]//IEEE International Conference on Multimedia and Expo. [S.l.: s.n.], 2005: 864–867.
- [68] Fu L, Mao X, Chen L. Speaker independent emotion recognition based on svm/hmms fusion system[C]//International Conference on Audio, Language and Image Processing. [S.l.: s.n.], 2008: 61–65.
- [69] Schuller B. Towards intuitive speech interaction by the integration of emotional aspects[J]. Proceedings of the Smc Yasmine Hammamet, 2002, 6: 6 pp. vol.6.
- [70] Kim E H, Hyun K H, Kim S H, et al. Speech emotion recognition using eigen-fft in clean and noisy environments[C]//The IEEE International Symposium on Robot and Human Interactive Communication, 2007. Ro-Man. [S.l.: s.n.], 2007: 689–694.

- [71] Zhou J, Wang G, Yang Y, et al. Speech emotion recognition based on rough set and svm[C]// IEEE International Conference on Cognitive Informatics. [S.l.: s.n.], 2006: 4898–4901 Vol. 8.
- [72] Pereira C. Dimensions of emotional meaning in speech[M]. [S.l.: s.n.], 2000.
- [73] Williams C E, Stevens K N. Vocal correlates of emotional state[M]. [S.l.: s.n.], 1981.
- [74] Johnstone T, Scherer K R. Vocal communication of emotion[M]. [S.l.: s.n.], 2000.
- [75] Cowie R, Cornelius R R. Describing the emotional states that are expressed in speech[J]. *Speech Communication*, 2003, 40(1): 5–32.
- [76] Ackroyd M H. Digital processing of speech signals[J]. *Electronics & Power*, 1997, 25(4): 290.
- [77] Cahn J E. The generation of affect in synthesized speech[J]. *Journal of the American Voice I/o Society*, 1990, 8: 1–19.
- [78] Proakis J G, Hansen J H. Discrete time processing of speech signals[M]. [S.l.]: Macmillan Pub. Co. ;, 1993
- [79] Gobl C. The role of voice quality in communicating emotion, mood and attitude[M]. [S.l.]: Elsevier Science Publishers B. V., 2003: 189–212
- [80] Lugger M, Yang B. The relevance of voice quality features in speaker independent emotion recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.: s.n.], 2007: IV–17 – IV–20.
- [81] Lugger M, Yang B. Psychological motivated multi-stage emotion classification exploiting voice quality features[M]. [S.l.]: InTech, 2008
- [82] Li X, Tao J, Johnson M T, et al. Stress and emotion classification using jitter and shimmer features[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.: s.n.], 2007: IV–1081–IV–1084.
- [83] Dhillon I S, Mallela S, Kumar R. A divisive information theoretic feature clustering algorithm for text classification[M]. [S.l.]: JMLR.org, 2003: 1265–1287
- [84] Dietterich T G. Approximate statistical tests for comparing supervised classification learning algorithms[M]. [S.l.]: MIT Press, 1998: 1895
- [85] Verweridis D, Kotropoulos C. Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition[J]. *Signal Processing*, 2008, 88 (12): 2956–2970.
- [86] Yang J, Honavar V. Feature subset selection using a genetic algorithm[M]. [S.l.]: Springer US, 1998: 44 – 49
- [87] Ben-Hur A, Guyon I. Detecting stable clusters using principal component analysis[J]. *Methods in Molecular Biology*, 2003, 224(224): 159.
- [88] Blum A L, Langley P. Selection of relevant features and examples in machine[J]. *Artificial Intelligence*, 1997, 97(1–2): 245–271.
- [89] Dempster A P. Maximum likelihood from incomplete data via the em algorithm[J]. *Journal of the Royal Statistical Society*, 1977, 39.
- [90] Burges C J C. A tutorial on support vector machines for pattern recognition[M]. [S.l.]: Kluwer Academic Publishers, 1998: 121–167

- [91] Kim Y, Lee H, Provost E M. Deep learning for robust feature generation in audiovisual emotion recognition[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. [S.l.]: IEEE, 2013: 3687–3691.
- [92] Deng J, Zhang Z, Marchi E, et al. Sparse autoencoder-based feature transfer learning for speech emotion recognition[C]//Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. [S.l.]: IEEE, 2013: 511–516.
- [93] Lee C C, Mower E, Busso C, et al. Emotion recognition using a hierarchical binary decision tree approach[C]//INTERSPEECH. [S.l.: s.n.], 2009.
- [94] Liu Y, Yao X. Ensemble learning via negative correlation[J]. Neural Networks, 1999, 12(10): 1399–1404.
- [95] Liaw A, Wiener M, et al. Classification and regression by randomforest[J]. R news, 2002, 2(3): 18–22.
- [96] Rätsch G, Onoda T, Müller K R. Soft margins for adaboost[J]. Machine learning, 2001, 42(3): 287–320.
- [97] Weston J, Tipping M. Use of the zero norm with linear models and kernel methods[J]. Journal of Machine Learning Research, 2003, 3(2003): 1439–1461.
- [98] Forman G. An extensive empirical study of feature selection metrics for text classification[M]. [S.l.]: JMLR.org, 2003: 1289–1305
- [99] Bekkerman R, Ran E Y, Tishby N, et al. Distributional word clusters vs. words for text categorization[J]. Journal of Machine Learning Research, 2003, 3: 1183–1208.
- [100] Kohavi R, John G. Wrappers for feature selection[J]. Artificial Intelligence, 1997.
- [101] Torkkola K. Feature extraction by non parametric mutual information maximization[J]. Journal of Machine Learning Research, 2003, 3(3): 1415–1438.
- [102] Barker L. Pattern classification[J]. Pattern Analysis & Applications, 1998, 1(2): 142–143.
- [103] Jolliffe I T. Principal component analysis and factor analysis[M]//Principal component analysis. [S.l.]: Springer, 1986: 115–128
- [104] Golub G H, Reinsch C. Singular value decomposition and least squares solutions[J]. Numerische mathematik, 1970, 14(5): 403–420.
- [105] Mika S, Ratsch G, Weston J, et al. Fisher discriminant analysis with kernels[C]//Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop. [S.l.]: Ieee, 1999: 41–48.
- [106] Genkin A, Lewis D D, Madigan D. Large-scale bayesian logistic regression for text categorization[J]. Technometrics, 2007, 49(3): 291–304.
- [107] Dasarathy B V. Decision fusion: volume 1994[M]. [S.l.]: IEEE Computer Society Press Los Alamitos, CA, 1994
- [108] 薛瑞尼. THUThESIS: 清华大学学位论文模板[EB/OL]. 2017. <https://github.com/xueruini/thuthesis>.

## 致 谢

衷心感谢导师吴志勇老师对我的精心指导，也感谢香港中文大学的 Helen 老师在论文发表过程中为我提供的帮助，他们的言传身教将使我终生受益。

感谢课题组的同学们对我入学以来提供的帮助，也感谢实验室的其他同学在学术上给我提供的助力。

感谢清华大学为我提供了舒适的学习和生活环境，让我在美丽的校园中度过了难忘的三年时光。

感谢 L<sup>A</sup>T<sub>E</sub>X 和 THUTHESIS<sup>[108]</sup>，帮我的论文写作节省了不少时间。

## 声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名： \_\_\_\_\_ 日 期： \_\_\_\_\_

## 个人简历、在学期间发表的学术论文与研究成果

### 个人简历

1992 年 6 月 28 日出生于甘肃省天水市。

2010 年 9 月考入北京邮电大学计算机系网络工程专业，2014 年 7 月本科毕业并获得工学学士学位。

2015 年 8 月考入清华大学计算机科学与技术系攻读计算机工程硕士学位至今。

### 发表的学术论文

- [1] X. Ma, Z.Y. Wu, J. Jia, M.X. Xu, H. Meng and L.H. Cai “Speech Emotion Recognition with Emotion-Pair based Framework Considering Emotion Distribution Information in Dimensional Emotion Space,” Proc. Interspeech, 2017. (CCF C 类, EI 收录, 检索号: 20175204591394)
- [2] X. Ma, D. Wang, J. Tejedor “Similar Word Model for Unfrequent Word Enhancement in Speech Recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol 24, no. 10, 2016. (CCF B 类, SCI 收录, 检索号: WOS:000381442600012)
- [3] X. Ma, X. Wang, D. Wang and Z. Zhang, “Recognize foreign low-frequency words with similar pairs”, Proc. Interspeech, pp. 458-462, 2015. (CCF C 类, EI 收录, 检索号: 20160902029708)
- [4] X. Ma, X. Wang and D. Wang, “Low-frequency word enhancement with similar pairs in speech recognition”, Proc. IEEE China Summit & International Conference on Signal and Information Processing, pp. 343-347, 2015. (EI 收录, 检索号: 20160701912086)