上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

学士学位论文

THESIS OF BACHELOR

论文题目：语音富信息识别算法研究

学生姓名：牛健威

学生学号：5100519066

专　　业：计算机科学与技术专业

指导教师：钱彦旻

学院(系)：电子信息与电气工程学院

语音富信息识别算法研究

摘要

语音富信息是指一段语音中所包含的关于说话人的信息，如情感状态、年龄、性别或者口音等信息。利用好这些语音富信息可以在人机交互中给用户更好的用户体验，系统也可以针对不同的人群提供不同服务。以隐马尔可夫模型（HMM）为基础，高斯混合模型（GMM）作为分类器的语音富信息识别系统在近些年得到了比较成功的应用。本文在此基础上，使用了一种深度神经网络模型代替 GMM 分类器在识别系统中的角色。这种深度神经网络的训练分为初始化和网络权值微调两个步骤。深度神经网络的网络权值是由多层的受限玻尔兹曼机（RBM）层层叠加形成的深度信念网络（DBN）来初始化，实践证明使用这样的算法初始化后的神经网络在训练时可以避免陷入局部最优解，从而大大提高了神经网络的学习和建模能力；之后网络权值再使用误差反向传播算法进行微调使得网络权值达到最优。本文使用了不同的语音特征作为神经网络的输入值，并且在此基础上融合了不同语音特征来提高识别率。实验结果表明，使用深度神经网络可以使情感识别的准确率相比 GMM 基线提高 8.0%，年龄识别准确率提高 5.7%，对不同语音特征进行融合后，又可以使情感识别准确率再提高 0.2%。

关键词：情感识别，高斯混合模型，深度神经网络，受限玻尔兹曼机，深度信念网络

STUDY OF RICH INFORMATION RECOGNITION IN SPEECH

ABSTRACT

Rich information in a speech utterance like the emotion state, age, gender or accent of a person has been intensively studied these years. Traditionally the recognition of these information has been using Gaussian Mixture Models (GMMs) for classification. However, the Gaussian Mixture Models do not make good use of multiple frames of input data and cannot exploit the high dimensional dependencies of feature sets efficiently, thus it's hard to improve the recognition accuracy for achieving a better result. In this paper, we introduce a deep neural network (DNN) that is first initialized by pre-training multi layers of Restricted Boltzmann Machines (RBM) and then fine-tuned using back propagation algorithm. The well-trained deep neural networks are capable of modeling complex and non-linear features of input training data and can better predict the probability distribution over classification labels. Many previous works in this field showed that this kind of DNN worked well in practice. In our work, we tried to replace GMM classifiers by DNN in the recognition system and found that the acoustic emotion and age recognition rate were both improved significantly compared with GMM baselines: the emotion recognition rate improved by about 8.0 percentage points and age recognition rate improved by about 5.7 percentage points. We mainly investigated the emotion recognition and conducted a series of experiments to find out the best system architecture. Furthermore, we used combination of different acoustic features to improve the emotion recognition rate by another 0.2 percentage points.

Key words: emotion recognition, Gaussian mixture model, deep neural network, restricted Boltzmann machine, deep belief network

目　录

第一章 绪论

语音是人与人之间交流的最有效、便捷的方式之一。近些年来，随着计算能力的不断提升，以及诸多人工智能、机器学习等领域中理论的不断完善和发展、实验和研究，使用机器对人类语音中包含的各种信息进行识别正在成为一个大的发展方向。本章先简要介绍语音识别的发展，之后再着重介绍关于语音富信息的研究。

1.1 语音识别概述

语音识别技术研究最早在二十世纪五十年代的 AT&T 贝尔实验室中，Audry 系统是第一个可以识别十个英文数字的语音识别系统。在之后的几十年中，新的技术理论和实验成果不断被提出和实践。六十年代，动态时间规整 (Dynamic Time Warping, DWT) 算法被提出；七十年代，矢量量化 (Vector Quantization, VQ)、隐马尔可夫模型 (Hidden Markov Model, HMM) 思想和模型等被提出，再加上线性预测编码 (Linear Prediction Coding) ，这些技术的发展，使得语音识别技术取得了很大进展；八十年代，HMM 模型得到了比较成熟的发展，也成为了语音识别技术中的重要和主流的模型，人工神经网络 (Artificial Neural Network, ANN) 被人们在语音识别领域应用起来。直到最近几年，随着计算能力的提高，深度神经网络模型正逐渐成为语音识别模型的主流方法之一，它已经被成功应用到音素识别、字词识别、情感识别等方面。

1.2 语音富信息研究介绍

在语音识别中，除了单词识别、语义识别、说话人识别等之外，语音富信息诸如情感状态、年龄、口音等额外的信息也逐渐被人们利用起来[1-5]。本文着重于语音富信息中情感状态的分类研究。为了获取说话人的情感，可以通过多种渠道收集各种信息和数据，如视觉方面的面部表情、手势，听觉方面的语音信号，甚至是一个人的血压、脉搏等生理信号。这样多方面的判断有助于情感识别率的提高，然而在某些以语音信号为主要沟通渠道的情况下，对说话人语音信号的分析和提取情感特征就显得尤为重要。

1.2.1 富信息情感识别研究现状

许多对语音情感分类的研究工作[6-14]都证明了对语音特征和语音情感分类模型的选择是影响情感识别准确率的两个重要的因素。Stuhlsatz A和Meyer C等人[6]使用了一种广义判别分析（GerDA）的特征提取和选择方法，其使用的语音特征表示值有过零率、基频、梅尔频谱参数和梅尔频率倒频谱系数 （Mel frequency cepstrum coefficients, MFCC） 等，并使用深度神经网络（Deep Neural Network, DNN）作为分类器，他们的实验结果表明使用深度神经网络对所提取的语音特征进行建模可以得到比用支持向量机作为分类器更出色的结果；Neiberg D等人[7]在瑞典Voice Provider公司的数据库中提取了语音信号中的MFCC参数和音高 （pitch） 参数，并使用不同频带的MFCC参数进行实验：滤波器分布在300Hz-8000Hz范围内的MFCC和分布在20Hz-300Hz范围内的MFCC-low，之后使用高斯混合模型 （Gaussian Mixture Model, GMM）作为分类器，他们的实验表明MFCC参数相比pitch参数可以更好地描述训练数据的情感特征，并且特征融合之后的分类效果又高于单用MFCC的结果；Kim Y等人[8]使用了由限制玻尔兹曼机 （Restricted Boltzmann Machine, RBM） 堆成的深度信念网络（Deep Belief Network, DBN） 作为深度学习模型，并提取了语音信号中的音高 （pitch），能

量（energy）以及梅尔频率滤波器组（Mel-frequency filter banks, MFB）的输出参数作为语音特征，并结合了视觉特征来进行情感识别，他们的实验结果比基线使用支持向量机（Support Vector Machine, SVM）的结果好，表明了语音信号中的高维非线性特征在情感识别中起着很重要的作用。Le D和Provost E M在他们的论文[9]中使用MFCC作为语音特征，并提出了一种混合DNN-HMM分类器，其中原来GMM的输出被DNN所代替。他们的实验在FAU Aibo数据库中进行训练和测试并取得了最好的实验结果。这一系列的实验结果可以看出，对语音情感识中特征的提取和分类器的选择是两个重要的方向，而对于特征的选择主要是一些韵律特征，如音高、基频、能量，以及一些频域特征如MFCC，MFB等。另外，选择不同的情感数据库做特征提取和分类器训练会得到不同的结果，这是由于不同情感数据库语料获取方式的差异造成的。如一种采集情感语音的方法是让演员在特定的环境下，想象自己处于那个环境再朗读出来；而一些情感语音是在特定场合引导出来的，如儿童语音数据库FAU Aibo[15]，是年龄在10-13岁的孩子在游戏过程中的自然录音，工作人员让他们相信游戏系统可以根据他们的口头指令做出相应的执行，而实际上该游戏系统是由工作人员在暗中操控的。

情感识别研究是在对语音特征参数的分析基础上对未知的语音信号进行情感识别。让机器对未知的语音信号进行情感识别是一件比较困难的事情，因为就算是人类有时也不一定对一句话中包含的情感能准确地区分开来，而且人类对情感的识别是建立在许多的基础之上并且考虑了多方面因素，比如说话人的面部表情、肢体动作，当时的语言上下文环境，语境等等。而这些人类对情感的识别是建立在自己熟悉的语音基础上，人类对自己不熟悉的语言的情感识别能力还要差一些。近些年许多研究者使用不同的方法让机器能识别语音信号中的情感，如上述文献所做的工作，并且取得了不错的成果。

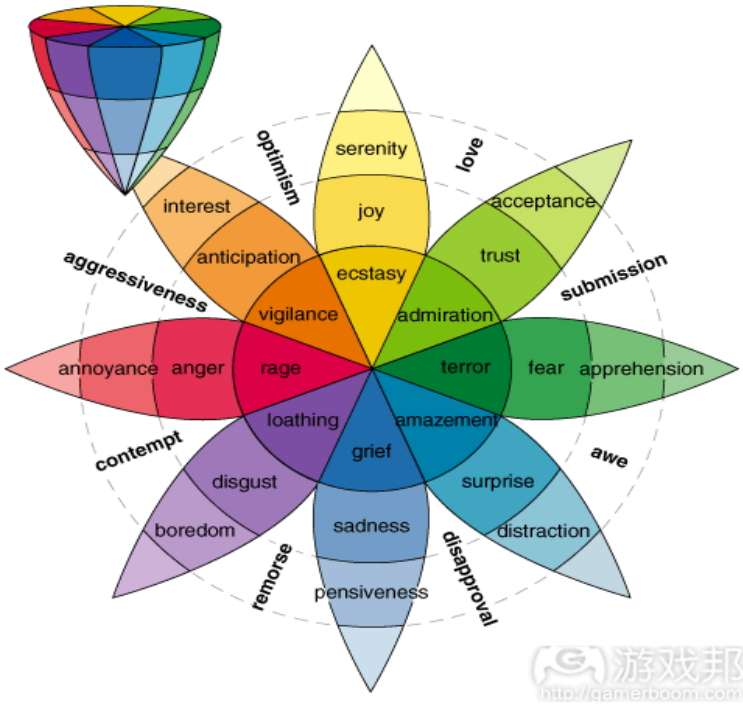

图 1-1 Plutchik 情感模型

1.2.2 情感描述模型

对情感的描述有离散和维度两种模型。

离散情感模型是指对情感的描述是以标签形式来刻画的，这种方式没有一个统一的定量或测量标准，一开始人们是使用日常生活中比较常用的词语来描述情感状态，如恐惧、悲伤、

高兴、惊讶等，但是毕竟基本的离散的情感状态的描述能力有限，后来有不同的学者通过对情感状态的分析提出了不同的情感模型，如在心理学领域比较知名的Robert Plutchik教授提出的八种基本情感状态、四个组合模型[16]（见图1-1）：anger-fear, anticipation-surprise, joy-sadness, trust-disgust。

每种不同的情感代表不同的颜色，而不同的颜色的叠加可以生成新的情感。另外，主要的一些学者及其主张的情感分类可以见表1-1。

表1-1 不同学者对情感的分类

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
James	Fear, grief, love, rage
McDougall	Fear, disgust, elation, fear, subjection, tender-emotion, wonder
Mower	Pain, pleasure
Oatley, Johnson-Laird	Anger, disgust, anxiety, happiness, sadness
Panksepp	Anger, disgust, anxiety, happiness, sadness
Plutchik	Anger, anticipation, disgust, joy, fear, sadness, surprise, trust
Tomkins	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Watson	Fear, love, rage
Weiner, Graham	Happiness, sadness

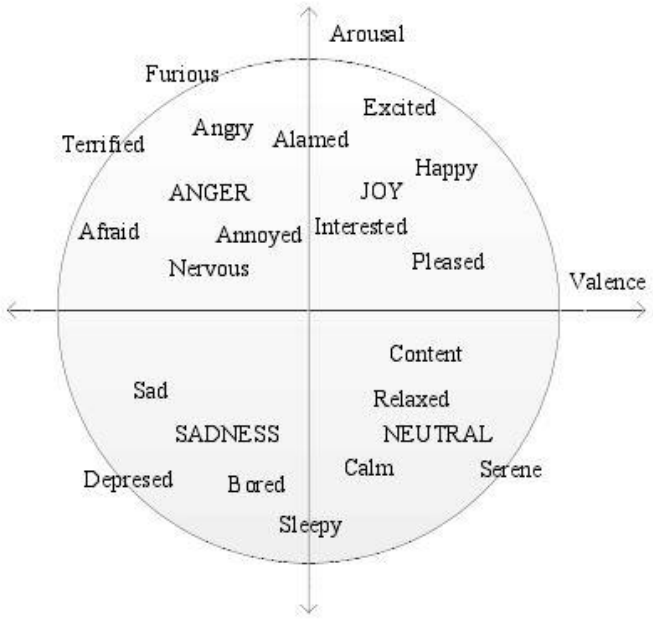

图 1-2 一种 Valence-Arousal 情感空间

从上述的表格来看，不同的学者对情感的分类还是有比较大的分歧，当然，其中都有某

些共性，讲这些分类的相似之处进行合并后，大概可以分类出几种基本的情感如愤怒（anger）、厌恶（disgust）、高兴（joy）、悲伤（sadness）、惊奇（surprise）等几大类。本课题使用的是离散情感模型，但我们的情感标签与这些基本的情感有共同之处也有一些变化，具体的情感分类参见第四章中的实验描述。

现在大多数的情感识别研究都是建立在离散情感模型基础之上，然而除了离散的情感模型之外，一些情感识别的工作使用的是另一种描述情感的方法：维度模型。维度模型理论中比较著名的是由三个认知维度组成的情感空间：Valence-Arousal-Power。简单来说，Valence维度主要描述情感状态的正负反馈，如是正面的（Positive）还是负面的（Negative）；Arousal维度描述某种正面或负面反馈的激励程度；Power维度表示人对情感状态的主观控制程度，如是主动的还是被动的。为了简化，一些工作只采用了前两个维度来对情感进行描述和研究，如图1-2。从图中可以看出，Valence值和Arousal值都比较大的，也就是出现正反馈并且比较强烈的情感都是一些比较积极的情感状态，如愉悦（joy）、高兴（happy, pleased）、激动（excited）等等；而Valence值和Arousal值都是负值的，也就是出现负反馈并且此负反馈不是很强烈的情感都是比较消极的情感状态，比如伤心（sad）、沮丧（depressed）、无聊烦人的（bored）等等；在图的第二象限是那些Valence值为负且Arousal值为正的情感状态，它们都表现出很强烈的负面情感，如生气（angry）、恐惧（terrified）、紧张（nervous）等；图的第四象限是那些Valence值为正但Arousal值为负的情感状态，它们大多表现出不强烈的正面情感，如舒服（relax）、平静（calm）等等。

1.2.3 语音富信息识别的应用场景

语音富信息包括了情感状态、年龄、口音、性别等信息，随着各种富信息识别研究的不断发展，将来定会给人们的生活带来显著的影响。就本文重点研究的情感分类来说，它的应用可以有以下几个方面：

（1）娱乐产业。现在市面上已经有可以进行情感交互的玩具，这种玩具可以对玩家的语音情感做出适当的回应。比如嵌入了情感识别系统的玩具狗，会根据主人对其的赞赏或责骂后表现出不同的喜怒哀乐的状态。

（2）工业。以后可能的一个发展方向是智能车载系统。当智能车载系统能“理解”车主的口令并做出相应的动作时，如果能够再进行语音的情感分析，则可以人性化地为车主提供不同的服务，比如当路况较差、车主心情烦躁的时候可以适当地提醒车主集中精力或者安抚心情，同时也提高了路面行车的安全性。

（3）服务业。在将来的客户服务中心，系统会根据实际情况来判断客户的情感状态，而为其选择一条合适的服务通道。这样不仅可以充分利用资源，也可以避免一些不必要的麻烦。另外，情感交互的系统也可以提高客户服务中心的服务质量：根据接线员的情感状态来进行不同的反馈如鼓励、安慰或赞扬等。

（4）医疗行业。当一些心理疾病患者就医时，医生有时不会发现一些细微的情感变化，或者不能很好地与病人进行沟通，从而耽误了治疗。情感识别系统可以给患者提供一个单独的环境来和患者交流，从而发现患者的心理状态和情感状态，以做出相应的动作，如鼓励和安慰等等，而这些数据也可以为医护人员提供一些参考价值，来更好地为患者服务。

当然，如果再将年龄识别、性别识别、口音识别等考虑进来，则语音富信息识别系统还将在更多领域有很大的贡献和发展空间，这里就不再一一赘述了。

1.3 本文主要研究内容和章节安排

在本课题中，我们主要关注对富信息情感的识别研究，我们使用了深度神经网络和高斯混合模型的识别结果进行对比，并探究了使用不同语音特征参数对深度神经网络识别结果的

影响。此外，我们还进行了富信息年龄和方言的识别研究。

本文的结构如下：第一章为绪论，介绍了语音富信息情感的研究现状，情感描述的模型和富信息识别的应用场景；第二章主要介绍了语音富信息特征提取的合理性，以及不同语音特征的提取方法；第三章介绍隐马尔可夫模型和本课题使用的分类器模型，包括高斯混合模型和深度神经网络模型；第四章讨论了实验内容和实验结果；第五章总结全文。

第二章 语音特征的提取与分析

2.1 引言

人声的产生是由声门、声道、口腔、鼻腔等一系列器官共同完成的：由声门肌肉的张力和从肺部压迫出的空气流，使得声门快速打开和关闭，就形成了声音。之后人的口腔、鼻腔和声道会影响到声音的音色，声门的震动频率影响到声音的基频，而来自肺部空气的力量大小决定了音量的大小[17]。人能从语音信号中体会到情感信息，有一部分因为语音信号中包含了不同的韵律特征等声学特征。韵律学特征包括声音信号的音高、能量、基频、时长等信息，这些信息已经被认为是区分情感的有效的特征。然而韵律学特征也有它的不足之处，比如“惊喜”和“愤怒”的特征比较相近，“伤心”和“害怕”的特征比较相近，所以对于这些相似度比较高的情感状态时，韵律学特征的区分能力就十分有限了。本课题中使用的声学特征是基于频谱以及基于倒频谱的声学特征，如梅尔频率倒谱系数 (MFCC)，感知加权线性预测系数 (Perceptual Linear Predictive，PLP) 以及滤波器组参数 (Filter Banks, FBANK)。基于谱的声学特征是声道形状的变化和发生运动 (articulator movement) 之间相关性的体现，这类声学特征已经被成功地运用到了语音识别、说话人识别[18][19]等领域。而在情感识别等语音富信息应用方面，MFCC、PLP 等特征也被认为是能够描述情感信息的有效特征，并且也有诸多的实验对此进行了验证。

2.2 语音信号的预处理

一般语音特征提取的流程如图 2-1 所示。

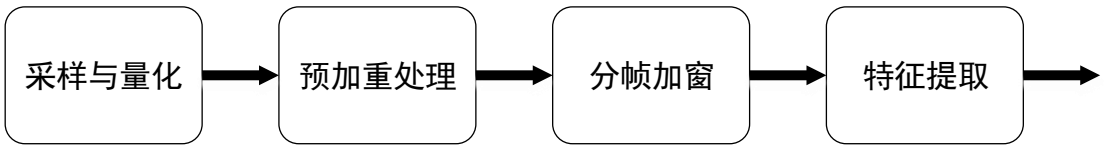

图 2-1 特征提取流程

由于现实生活中的声音是模拟量，所以先对声音信号进行采集，将其转化为离散的数字信号以便之后处理。通常录制的语音信号不能直接拿来进行特征提取，因为语音信号中存在着许多冗余信息，如果直接进行特征提取可能还会增加计算量并且降低识别率，所以要先对语音信号进行预加重处理。之后再进行分帧加窗得到一帧一帧的语音信号，这是提取特征的基本单位。最后再对这些经过处理的语音帧进行特征提取。

对语音信号的分析是基于短时分析技术。语音信号时刻都在发生变化，是一个非平稳过程，但若考虑语音信号在某个很短的时间内 (通常是 10ms 至 30ms)，则可以大致认为语音信号是平稳的。只有建立在这样的基础上，才能对每一帧的语音信号进行特征提取和分析。每一帧短时平稳的语音都会被做相同的处理、提取短时特征参数，之后一段原始的语音信号就可以用这些短时特征参数来表示，为之后的分类器的训练和测试做准备。

2.2.1 语音信号的预加重

语音信号的预加重通常是通过一个一阶高通滤波器：

$$H(z) = 1 - az^{-1} \tag{2-1}$$

其中通常取 a=0.97，是预加重系数。

由于连续的语音模拟信号在采样后变为离散信号，所以所有的数据类型都为离散类型。假设S(n)为预加重前的语音信号，n 为某一时刻的语音信号的编号，则预加重后的语音信号$\tilde{S}(n)$可以表示为

$$S(n)=S(n)-aS(n-1) \tag{2-2}$$

2.2.2 语音信号的分帧加窗

之前说过，语音信号是随时间变化的，但在某个短时间内可以看作是平稳变化的，一般这个短时时间段取 10ms-30ms，在本次实验中取 25ms。当把语音信号以 25ms 为一个处理单位后(即为一帧)，会为每帧信号做加窗处理。一般来说加窗时会与前后的帧有所重叠，这样可以保持平滑过度和连续型，以避免损失过多的信息。帧长和帧移的示意图如图 2-2。

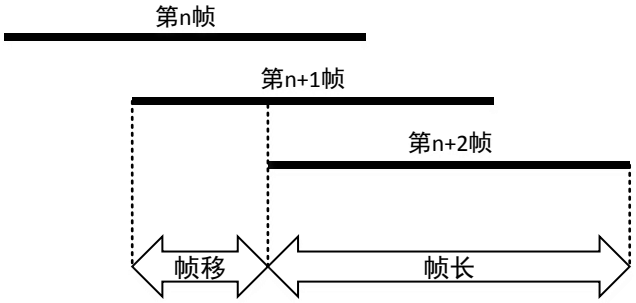

图 2-2 帧长和帧移的示例

之后用一个窗函数w(n)与$\tilde{S}(n)$相乘得到加窗后的信号S_w：

$$S_w=S(n)*w(n) \tag{2-3}$$

其中窗函数有几种选择，如矩形窗：

$$w(n)=\begin{cases}1, & 0\le n\le N-1\\ 0, & \text{其它}\end{cases} \tag{2-4}$$

汉明窗：

$$w(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0\le \mathrm{n}\le \mathrm{N}-1\\ 0, & \text{其它}\end{cases} \tag{2-5}$$

通常选择汉明窗为窗函数，这样可以保证一帧中的中间部分信息不会被丢失，并且一帧的两端被平滑过度，而产生类似与周期性的效果，为后面的离散傅立叶变换做准备。本课题采用的也是汉明窗。

2.2.3 端点检测

端点检测技术是为了在一段语音信号中去除背景噪音，从而准确定位有用语音信息的起始端点。端点检测是基于时域信号的短时平均能量和过零率这两个参数，根据这两个参数的阈值来进行判定。一般来说，有声语音的短时平均能量值较大，而背景噪音等无声语音信号的过零率较高。

某段语音的短时能量值可以如下计算：

$$\mathrm{E}=\sum_{m=0}^{N-1}x^{2}(m) \tag{2-6}$$

$x(m)$代表离散语音信号，N 为帧数。

短时过零率可以如下计算：

$$Z=\frac{1}{2}\sum_{m=0}^{N-1}\left|\mathrm{sgn}[x(m)]-\mathrm{sgn}[x(m-1)]\right| \tag{2-7}$$

其中 sgn[x]代表一种符号函数：

$$\mathrm{sgn}[x]=\begin{cases}1, & x>0\\ -1, & x\le 0\end{cases} \tag{2-8}$$

2.3 特征参数的提取

情感识别中，谱特征是较为关键的参数，而不同谱特征对情感的描述有所不同，为了探索哪种语音特征对情感的识别率较高，本课题使用了三种不同的谱特征参数作为语音信号的特征参数，它们分别是梅尔频率倒谱系数，感知线性预测参数和滤波器组参数。

2.3.1 梅尔频率倒谱系数（MFCC）

目前研究工作中使用比较广泛、对语音信号建模能力较好的就是 MFCC 参数。由于人的听觉系统是一套关于声音频率的非线性系统，所以它对不同频率的声音的感知度不同，一般来说人耳对 1000Hz 以下的频率的感知成近似的线性关系，对 1000Hz 以上的高频区域则不是线性关系而是近似于对数关系。MFCC 的提取流程是将频率非线性地映射到梅尔频率坐标中进行的，这样一来人耳对梅尔频率坐标中的频率就近似于线性关系，这样可以较好地模拟了人耳的听觉系统。MFCC 参数也在语音识别、说话人识别等领域得到了广泛地应用。梅尔频率与正常的频率之间的转换关系是：

$$f_{mel}=2595\lg\left(1+\frac{f}{700}\right) \tag{2-9}$$

其中f_{mel}代表梅尔频率。它们之间的关系图如图 2-3 所示。这样转换之后，原先人耳对频率感知的非线性关系在梅尔频率坐标系中就变成了线性的，这样可以将滤波器组均匀地分布在梅尔频率坐标系中提取参数。

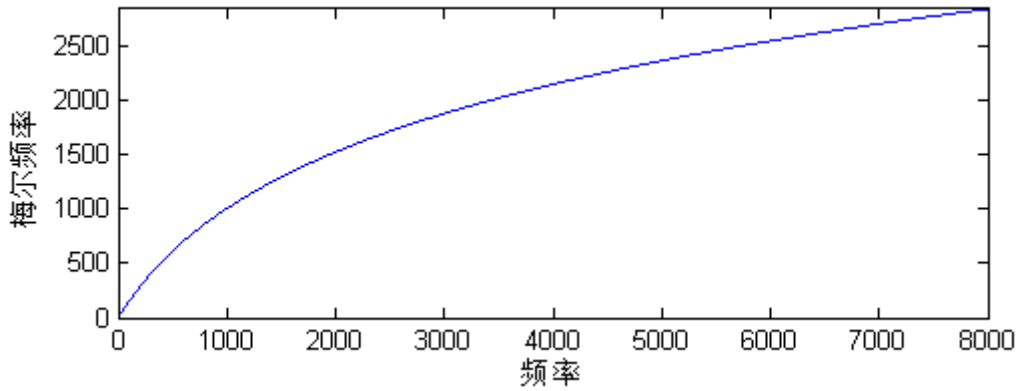

图 2-3 频率和梅尔频率关系转换示意图

特征参数的提取是针对每一帧进行的，步骤如下：

(1) 将经过预处理的每一帧语音$x(n)$做离散傅立叶变换，提取频谱参数：

$$X(k)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi k/N},\quad 0\le n,k\le N \tag{2-10}$$

N 代表 FFT 离散的点数；

(2) 对结果取模的平方得到离散功率谱；

(3) 将得到的频谱通过均匀排列在梅尔频率坐标轴中的三角滤波器组，得到滤波器组输出参数：

$$S(m)=\ln(\sum_{k=0}^{N-1}|X(k)|^2H_m(k)),\ 0\le \mathrm{m}\le \mathrm{K} \tag{2-11}$$

$H_m(k)$为三角滤波器组的频率响应函数，在梅尔频率坐标系中排列个数为 24 到 40 个之间，一般前 13 个滤波器在 1000Hz 以下，后 27 个滤波器高于 1000Hz。每个三角滤波器的相应函数为：

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(\mathrm{m}+1)-f(\mathrm{m}-1))(f(\mathrm{m})-f(\mathrm{m}-1))}, & f(m-1)\le k\le f(m)\\ \dfrac{2(f(\mathrm{m}+1)-k)}{(f(\mathrm{m}+1)-f(\mathrm{m}-1))(f(\mathrm{m}+1)-f(\mathrm{m}))}, & f(m)\le k\le f(m+1)\\ 0, & k\ge f(m+1)\end{cases} \tag{2-12}$$

(4) 将滤波器组的输出值先取对数，然后做相当于将频谱变为“伪频谱”的离散余弦变换，从而得到梅尔频率倒频谱系数，一般取前 12 个参数加上第零阶倒谱系数形成一个 13 维的向量，作为这一帧语音的特征参数：

$$C(n)=\sum_{m=0}^{N-1}S(m)\cos\left(\frac{n\pi(m-0.5)}{M}\right),\quad 0\le n\le K \tag{2-13}$$

这里C(n)为 MFCC 参数。MFCC 提取的流程图如图 2-4 所示。

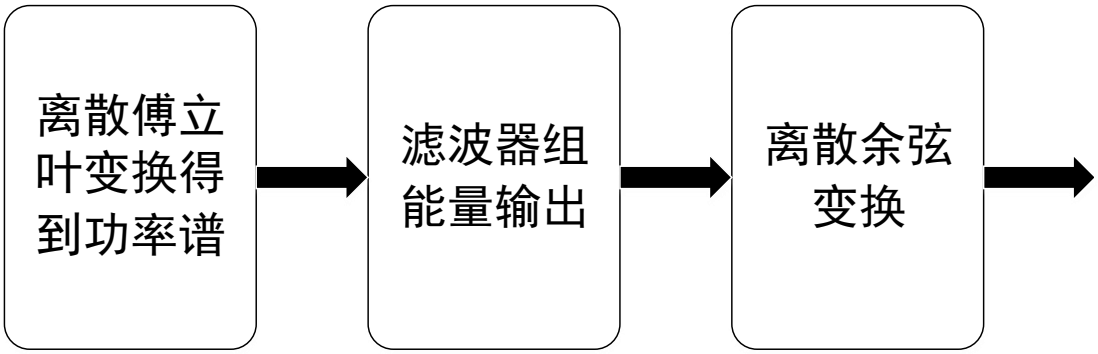

图 2.4 MFCC 提取流程

2.3.2 MFCC 的能量参数与动态参数

2.3.2.1 归一化能量

归一化能量和动态参数都是为了增加识别性能而加入的。语音信号的归一化能量是一个

一维分量，计算公式如下：

$$\mathrm{E}_i = \log \sum_{n=1}^{N} s_i^2(n) \tag{2-14}$$

其中 N 代表语音中的帧数。之后再对能量做归一化使得与提取的动态参数有可比性，以便一同运算：

$$\mathrm{E}_i = \mathrm{E}_i - \mathrm{E}_{\max} \tag{2-15}$$

其中 $\mathrm{E}_{\max}$ 为语音段中最大的能量值。

2.3.2.2 动态差分参数

在语音特征中加入动态差分参数可以是识别率有很大提高。本课题中一阶动态差分参数的计算公式为[20]：

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(\mathrm{c}_{t+\theta} - \mathrm{c}_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2-16}$$

这里d_t表示 t 时刻的动态差分参数，它是基于基础的静态参数$c_{t+\theta}$和$c_{t-\theta}$计算出来的。Θ是窗移大小，本课题中为 2。与计算一阶差分动态参数类似，二阶动态差分参数是在一阶差分参数的基础上计算的。一般在语音特征参数后加入一阶和二阶差分参数来提高识别准确率。

2.3.3 感知线性预测参数（PLP）

PLP 相比传统的线性预测参数 (LP) 考虑了更多的听觉和心理影响，在对语音建模能力更优于传统的 LP 参数，是一种基于听觉模型的特征参数[21]。PLP 参数的提取过程如下：

(1) 使用离散傅立叶变换求的功率谱；

(2) 临界频带分析。临界频带的划分表示了人耳听觉的掩蔽效应，体现了人耳的听觉模型。一个临界带宽单位用 Bark 来表示。临界频带编号 Z 与频率 f 之间的关系如下公式所示，式子中将频率 f 映射到 Bark 频率 Z，一般取$0 \le \mathrm{Z} \le 21$，去除首尾的两个频带后剩下 20 个频带：

$$Z = 6 * \ln\left(\frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1} \right) \tag{2-17}$$

(3) 每个频带内的能量谱与如下的加权函数相乘求和，得到听觉谱：

$$C_k(Z) = \begin{cases} 10^{Z - Z_k + 0.5}, & Z \le Z_k - 0.5 \\ 1, & Z_k - 0.5 \le Z \le Z_k + 0.5 \\ 10^{-2.5(Z - Z_k - 0.5)}, & Z \ge Z_k + 0.5 \end{cases} \tag{2-18}$$

(4) 等响度预加重。在相同声音强度下，人耳对不同频率的声音所感受到的强度是不同的，为了模拟人耳的听觉特性，对上式的输出做 lg40 dB 的等响度曲线变换。等响度曲线函数的公式为：

$$e(\omega) = \frac{\omega^2 * \left(\omega^2 + 1.44 * 10^6\right)}{\left(\omega^2 + 1.6 * 10^5\right) * \left(\omega^2 + 9.61 * 10^6\right)} \tag{2-19}$$

(5) 离散傅立叶反变换。将 20 点经过离散傅立叶反变换后，用德宾算法计算 M 阶全极点模型 (M 一般取 5-15)，并求出倒谱系数，最后的结果就是 PLP 参数。

提取 PLP 参数的流程如图 2-5：

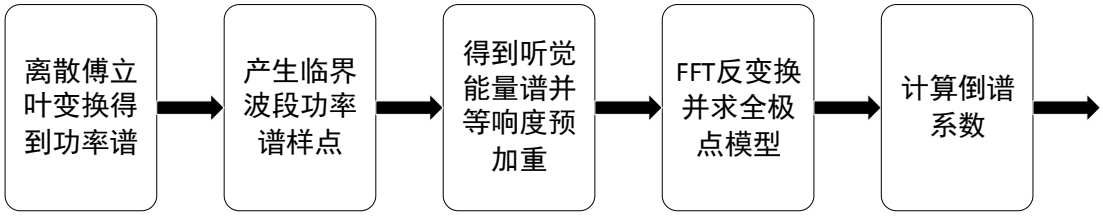

图 2-5 PLP 参数提取流程

2.3.4 滤波器组参数（FBANK）

FBANK 参数是直接基于滤波器组的输出，如图 2-6 所示。三角滤波器的中心频率均匀排列在梅尔频率坐标系中，三角滤波器的输出便是一种更直接的非线性的表示，因此许多时候也可以使用 FBANK 来代替线性预测参数。对滤波器组的输出 MELSPEC 取对数，则最终得到的就是 FBANK 参数，其中 FBANK 参数的维度与滤波器的个数有关，一般为 20 到 40 个，我们使用的具体的实验配置在第四章会介绍。在本课题中，还使用了特征的动态一阶差分项和二阶差分项来提高特征参数对语音信号的建模能力和最终的识别率。

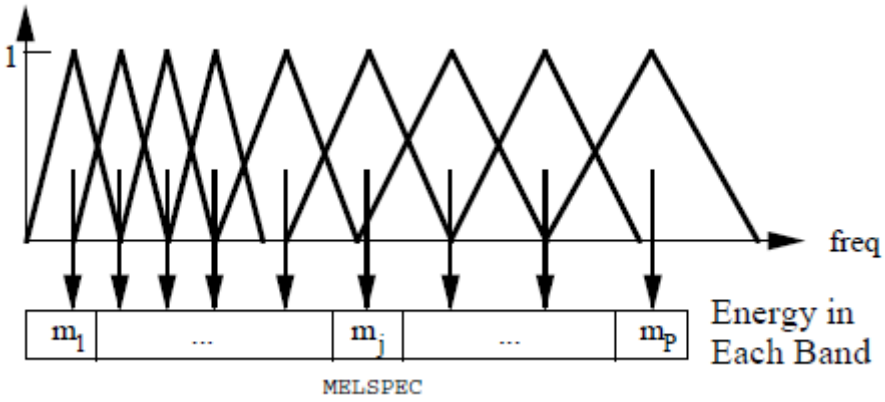

图 2-6 基于梅尔频率的滤波器组

2.4 本章小结

本章主要介绍了对于语音富信息的特征提取的合理性以及语音特征提取的流程。将输入的语音信号通过预处理后便可以根据需求提取不同的特征参数如 MFCC, PLP, FBANK，以及动态差分项的计算。之后的过程中，所有的语音信号都将以特征参数的形式体现，这为之后分类器的训练和测试做好了准备。

第三章 分类器模型

3.1 引言

在语音识别技术中，隐马尔可夫模型[22](HMM) 的成功应用使得识别技术有了很大的提高。近些年来已经有很多有关语音识别的实验应用 HMM 对语音序列进行建模并取得了很大的成功。本课题对语音富信息的建模也是建立在 HMM 之上，并分别使用了两种不同的分类器模型：高斯混合模型 (GMM) 和深度神经网络模型 (DNN)。

3.2 隐马尔可夫模型（HMM）介绍

HMM 是在马尔可夫链模型的基础上发展出来的。HMM 是一个有限状态机，它在某一个时刻的某一个状态根据一个概率模型跳转到另一个状态。HMM 比马尔可夫链更加复杂，它描述的是一个双重随机过程：一个随机过程描述状态的转移，另一个随机过程描述某一状态和观察值的统计对应关系。人的语音的产生过程与这个相似：它是由人的大脑中的不可见的根据语言知识发出的参数流。所以一个左右无跳转的 HMM 模型可以比较方便有效地模拟人的语音流，成为了对语音建模的不二选择。HMM 包括三种基本类型：离散型 (Discrete HMM)、连续型 (Continues HMM) 和半连续型 (Semi-continues HMM)。本课题使用的是连续型 HMM，它的每个状态的输出概率是由 GMM 的连续的概率密度函数所决定。

描述一个 HMM 通常可以用五个参数$\lambda = \{\pi, A, B, N, M\}$。其中π是描述起始状态分布概论的参数；A 是状态转移矩阵，也就是从某一时刻q_t处于状态S_i，跳转到q_{t+1}时刻处于状态S_j的概率，HMM 假设每一个时刻的状态之于之前一个时刻的状态有关；B 是在某个状态下的输出概率分布，若其概率分布为离散型则此 HMM 为离散型，若其概率分布为连续型则此 HMM 为连续型；N 是 HMM 模型中状态S_i的个数，在对语音识别的建模中常用的 N 的数目是 5，其中首末两个状态为开始和结束状态，它们只是为了建模的方便而加入，从而并没有输出概率分布；M 表示不同的观测事件的数目。图 3-1 是一个比较典型的五状态，左右无跳转的 HMM 模型，其中状态S_2，S_3和S_4是具有发射概率密度函数的状态，S_1和S_5是开始和结束状态。$O = (o_1, o_2 \dots o_\tau)$是观察值序列，$b_i(o_j)$表示在状态$S_i$下观测到值为$o_j$的概率。

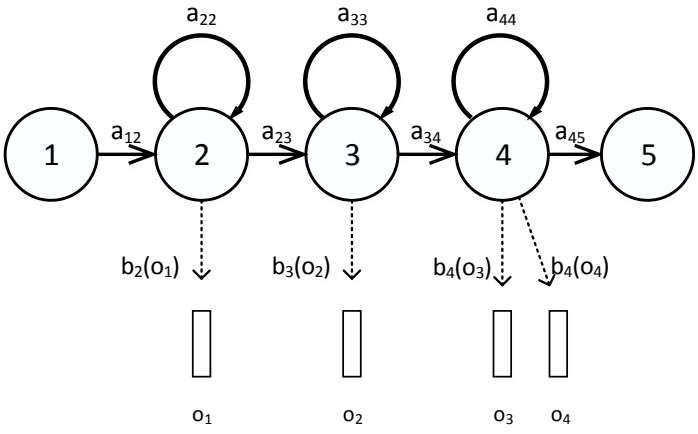

图 3-1 隐马尔可夫模型示例

更形象地说，HMM 的双重随机过程可以如图 3-2 所示：由参数(π,A)描述的产生观察序列的过程，以及由参数(B)描述的产生观察值的过程。

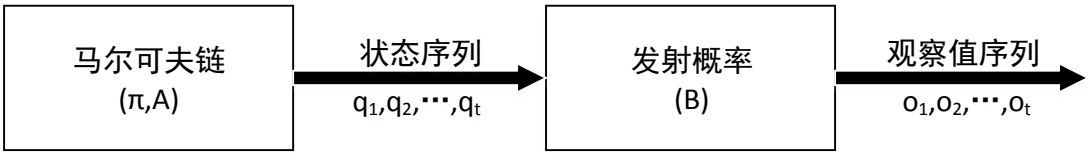

图 3-2 HMM 随机过程组成示意图

确定了 HMM 的参数，就可以得到一个确定的 HMM 模型。要使用 HMM 对语音序列进行建模，还需要解决 HMM 的三大经典问题：

(1) 评估问题：对于给定的一个模型$\lambda=\{\pi,A,B,N.M\}$和一个已知的观察序列

$O=(o_1,o_2,\ldots,o_\tau)$，求这个观察序列在此模型下出现的概率$P(O|\lambda)$，可以使用前向后向算法；

(2) 解码问题：对于给定的一个模型$\lambda=\{\pi,A,B,N.M\}$和一个已知的观察序列

$O=(o_1,o_2,\ldots,o_\tau)$，求一个状态序列$Q=\{q_1,q_2,\ldots,q_t\}$使得该状态序列出现的可能性最大，可以使用 Viterbi 算法；

(3) 学习问题：给定一个观察序列$O=(o_1,o_2,\ldots,o_\tau)$，调整模型参数$\lambda=\{\pi,A,B,N.M\}$使得该观察序列在此模型下出现的概率最大，可以使用 Baum-Welch 算法。

3.3 GMM-HMM 分类器

3.3.1 高斯混合模型（GMM）

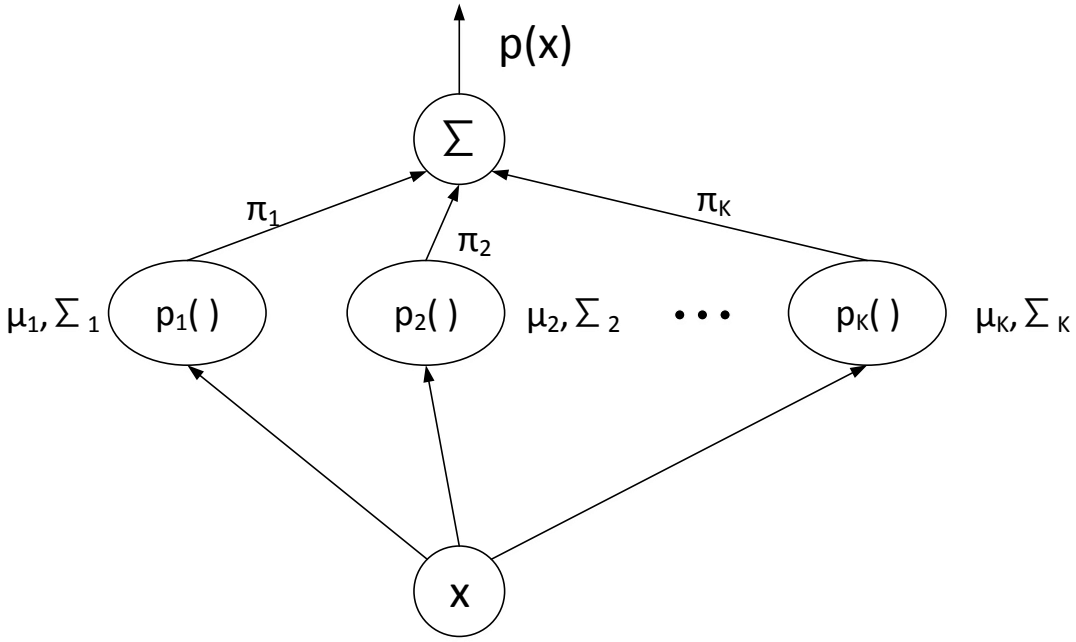

图 3-3 GMM 模型示意图

GMM 已经在语音识别的领域得到了很广泛的应用。GMM 是由多个高斯概率密度函数的线性组合而成的概率分布，从中心极限定理的角度来看，将未知的数据分布假设为高斯分布比较合适。GMM 的概率密度函数可以表示为：

$$p(x)=\sum_{k=1}^{K}\pi_k\,p(x\mid k) \tag{3-1}$$

其中 K 是高斯模型的个数，π_k为第 k 个高斯模型的权重，$p(x|k)$为第 k 个高斯模型，其均值为μ_k、方差为Σ_k的概率密度函数，示意图如图 3-3 所示。当 K 的值趋近于很大时，则可以认为 GMM 可以逼近任意的概率分布。对 GMM 进行训练时，使用 EM 算法求出当某个实例出现的概率最大时的系统参数，也就是使用最大化似然函数的方法使得下面的式子达到最大：

$$\max \quad \sum_{i=1}^{N}\log(\sum_{k=1}^{K}\pi_k N(x_i \mid \mu_k, \Sigma_k)) \tag{3-2}$$

由于许多概率的线性加权值通常是一个很小的实数，所以一般对结果取对数再进行之后的运算。以 HMM 为底层模型，GMM 为状态发射概率，就得到了一个可以对语音特征进行训练和分类的模型。

3.3.2 GMM-HMM 分类器

GMM-HMM 分类器多用于对音素或单词的建模，在本实验中使用它对富信息状态进行建模。一般来说，每个 HMM 状态会对应一个单词或音素，或者多个 HMM 状态对应一个单词或音素。每个 HMM 状态都会由一个 GMM 模型来计算它的发射概率，这个 GMM 就是对某个音素和单词所建立的模型，它在训练阶段使用 EM 算法调整参数来对音素或单词进行建模。HMM 在训练阶段使用 Baum-Welch 算法对状态转移矩阵等参数进行调节，以达到最好地描述某个音素或单词的在语音特征层次中的产生过程。

在解码阶段，使用 Viterbi 算法求出 HMM 状态转移的最佳序列，再对状态序列中的每个状态使用对应的 GMM 模型来求出当前时刻的发射概率。最终取概率和最大的那个 HMM 模型作为结果。在本课题中，仅将音素或单词的概念换成了富信息状态，所以使用的训练和识别方法是和传统的音素和单词是一样的。

3.4 NN-HMM 模型

3.4.1 人工神经网络（ANN）模型

人工神经网络[23]是一种旨在模仿人脑结构及其功能的信息处理系统[24]，它由大量的节点 (或者称为神经元) 相互连接构成。人工神经元模型从上世纪 40 年代被提出，经过了四十年的坎坷发展，到 80 年代由美国物理学家 Hopfield 提出了 Hopfield 模型[25]，从而使 ANN 得到大力的发展和应用。ANN 主要有以下几个特征：

(1) 分布式的信息存储。
(2) 具有联想功能和记忆功能。
(3) 具有较强的学习能力。
(4) 具有良好的高容错性和高度非线性。
(5) 具有大规模的自组织、自适应能力。

ANN 模型已经被成功应用到了信息处理、模式识别、专家系统、风险评估[26][27]等许多领域中。

3.4.1.1 神经元模型

一个典型的神经元模型是具有多个输入、单个输出的非线性结构，是组成庞大神经网络的基本单元，如图 3-4 所示。

其中$x_i(i = 1,2 \ldots m)$为输入向量，$w_{ij}(\mathrm{i} = 1,2 \ldots \mathrm{m})$为神经元之间的连接权重，$y_j$是神经元的输出。一般来说，输入向量在求加权和之后会通过一个非线性的激活函数f，再将激活函数的结果作为整个神经元的输出。所以上图可以用公式表示为：

$$s_j = \sum_{i=0}^{m} w_{ij} x_i \tag{3-3}$$

$$y_j = f(s_j) \tag{3-4}$$

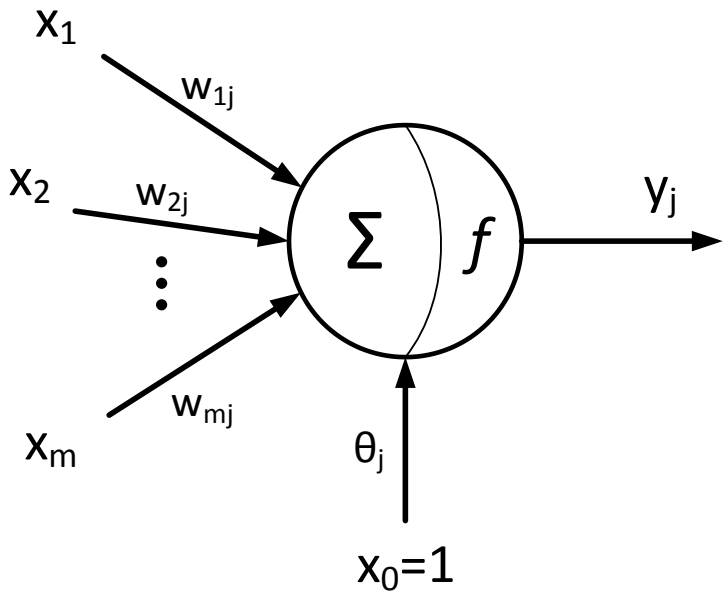

图 3-4 人工神经元示意图

激活函数$f(x)$可以根据不同的需求而选择不同的函数，比较常见的有：

(1) 线性函数：

$$f(x) = x \tag{3-5}$$

(2) 双曲正切函数：

$$f(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{3-6}$$

(3) 高斯核函数：

$$f(x) = \exp(-\frac{1}{2\sigma_i^2} \sum_j (x_j - w_{ji})^2) \tag{3-7}$$

(4) Sigmoid 函数：

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{3-8}$$

等等。本课题中选择的是 Sigmoid 函数。Sigmoid 函数将所有的输入挤压到 0-1 之间，并且它具有良好的数学性质如连续、可导等。Sigmoid 函数的形状如图 3-5 所示。

3.4.1.2 神经网络模型结构和 BP 算法

人工神经网络可以根据网络内神经元的连接方式不同而组成不同的结构。比较常见的有反馈式前馈神经网络、层内互联式神经网络、前馈式神经网络和相互连接式神经网络等。目前采用较多的是层内无连接的前馈式神经网络模型，当神经网络由输入层、中间层和输出层

组成时，就形成了一个多层感知器 (Multi-Layer Perceptron, MLP)，如图 3-6 所示。

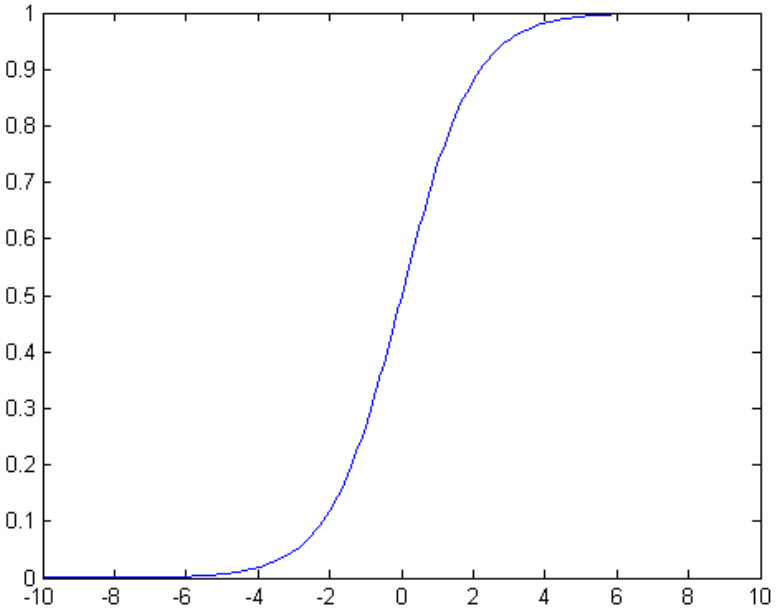

图 3-5 Sigmoid 函数示意图

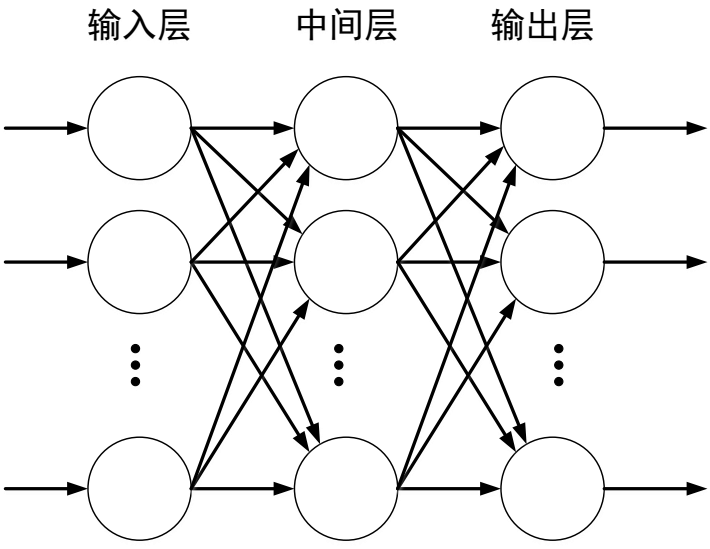

图 3-6 三层前馈神经网络模型

理论可以证明这样的三层 MLP 可以实现任意从输入到输出的映射关系。在训练前馈式的神经网络模型时，多用误差反向传播算法进行网络权值的调整。以三层前馈式神经网络为例，误差反向传播 (Back Propagation, BP) 算法的主要步骤如下：

(1) 网络权值初始化。给网络中的连接权值和阈值设定为[-1,1]之间的随机数；

(2) 随机选取一对样本和对应的输出： $\mathrm{x(k)} = (x_1(\mathrm{k}), x_2(\mathrm{k}), \ldots, x_n(\mathrm{k}))$ 和 $\mathrm{d(k)} = (d_1(\mathrm{k}), d_2(\mathrm{k}), \ldots, d_q(\mathrm{k}))$

(3) 计算中间层节点的输出值；

(4) 计算输出层节点的输出值；

(5) 计算输出层的实际值与期望输出值的误差，并计算该误差对输出层所有节点的偏导数；

(6) 计算中间层的输出的误差函数，并对中间层的各个节点求偏导数；

(7) 利用之前求出的偏导数对输出层的权值和阈值进行更新；

(8) 再对中间层的权值和阈值进行更新；

(9) 计算权值误差并判断误差是否满足要求，若误差满足要求并训练轮数达到了指定的轮数，则结束训练，否则返回步骤(2)。

BP 算法也存在这一些缺陷，如：

(1) BP 算法需要大量的已标注的数据，而很多时候大多数的数据都是未标注的；

(2) BP 算法的收敛时间比较难控制，尤其是当网络层数较多的时候；

(3) BP 算法容易陷入局部最优解。

这就需要取寻找更好的模型和算法进行求解。

3.4.2 深度神经网络（DNN）模型

随着计算硬件成本的降低和计算能力显著的提高，近些年来深度学习称为了语音识别、语音富信息识别的主要研究对象。本文将介绍一种深度信念网络 (Deep Belief Network, DBN) 模型，是一种概率生成模型。DBN 是由多层的受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM) 层层叠加而成，而 DBN 的训练过程也是逐层地进行网络权值和节点偏置的调整。这样做的好处是可以根据欲求任意地叠加 RBM，而计算的时间复杂度也和神经网络的深度成线性关系。DBN 的训练是属于无监督学习，可以用对大量无标注的数据进行训练，这对于在有标记数据量有限的情况下是一个很好的选择。另外，因为在网络隐层数很多的情况下，使用 BP 算法进行网络权值的调整很容易陷入局部最优解，而许多实验已经证明，使用训练好的 DBN 网络权值对 DNN 进行初始化后，再使用 BP 算法是可以有效地防止 BP 算法陷入局部最优解，从而使得效果比随机初始化要好得多。

3.4.2.1 受限玻尔兹曼机（RBM）模型

玻尔兹曼机是一种基于能量模型的随机递归神经网络模型。最先提出学习玻尔兹曼机的是 David Ackley 和 Geoffrey Hinton 在 1985 年发表的一篇论文[28]中。由于递归神经网络学习比较复杂，所以 Smolensky 等人提出了受限玻尔兹曼机模型[29]。受限玻尔兹曼机是将原先的玻尔兹曼机中加入了层内无连接这样的条件，形成了一个二部图，如图 3-7 所示，使得有方法可以快速地学习这样的神经网络。一个两层的 RBM 有以下几个特点：

(1) 层内无连接，层之间的节点全连接。

(2) RBM 是一个无向图。

(3) 当给定一个特定的可见层 (隐藏层) 输入时，隐藏层 (可见层) 的节点的激活状态是相互独立的。

图 3-7 显示的是一个两层的受限玻尔兹曼机，它由一层可见层和一层隐藏层组成。根据 RBM 中节点的取值不同有不同的类型，比较常见的是高斯型和二值型。在这里为了简单起见假设 RBM 中的节点取值为二值型：0 和 1。对于可见层向量和隐藏层向量的每一种可能的取值，有会有一个能量函数$E(v,h)$所对应。

$$E(\mathrm{v},\mathrm{h}\mid\theta)=-\sum_{i=1}^{V}a_i v_i-\sum_{j=1}^{H}b_j h_j-\sum_{i=1}^{V}\sum_{j=1}^{H}v_i w_{ij} h_j \tag{3-9}$$

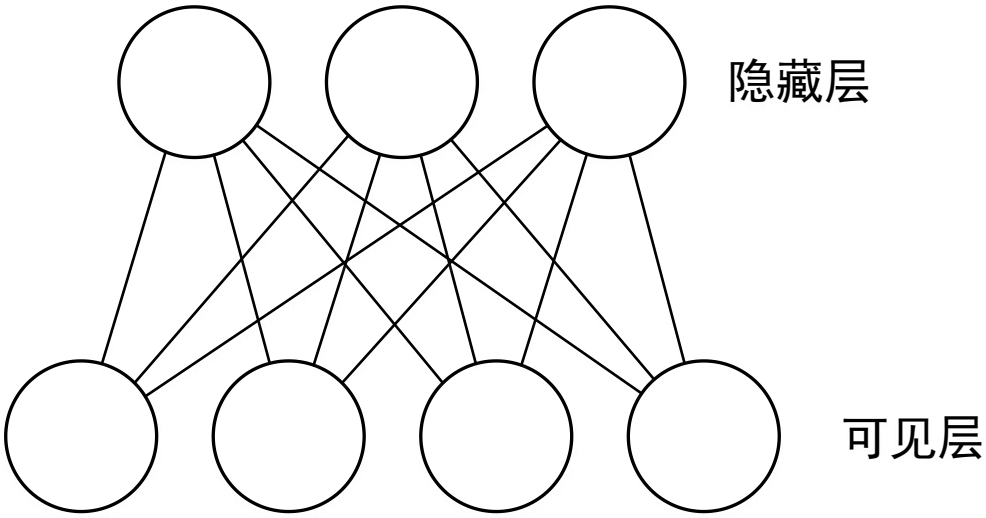

图 3-7 受限玻尔兹曼机模型

其中$\theta=\{w_{ij},a_i,b_j\}$是 RBM 的参数，$w_{ij}$表示节点之间的连接权重，$a_i,b_j$分别表示可见层节点和隐藏层节点的偏置。可见层和隐藏层的联合概率分布定义为：

$$P(\mathrm{v},\mathrm{h}\mid\theta)=\frac{1}{Z(\theta)}e^{-E(\mathrm{v},\mathrm{h}|\theta)} \tag{3-10}$$

$$Z(\theta)=\sum_{v,h}e^{-E(\mathrm{v},\mathrm{h}|\theta)} \tag{3-11}$$

其中Z(θ)是归一化参数 (partition function)，使得概率的和为 1。对 RBM 的训练是要使得关于可见层的边际概率分布最大，也即：

$$P(\mathrm{v}\mid\theta)=\frac{1}{Z(\theta)}\sum_{h}e^{-E(\mathrm{v},\mathrm{h}|\theta)} \tag{3-12}$$

为了得到上述分布，需要计算Z(θ)，但这个计算量是在2^{V+H}这个数量级，很显然在可见节点和隐藏节点数目很大的时候是无法有效计算出结果的。在 2006 年，Hinton 提出了一种 Contrastive Divergence (CD) 快速学习算法[30]。CD 算法使用了 Gibbs 采样得到 RBM 中可见层和隐层的分布，从了避免了原先为了得到隐层和可见层的稳步而需要的庞大的计算量。假设一个训练样本为x_0，隐层单元个数为 H，可见层单元个数为 V，学习率为ϵ，最大学习周期为 T，σ(x)是 Sigmoid 激活函数，则 CD 算法所采用的计算步骤为：

(1) 将可见层用样本初始化$v_1=x_0$，网络权值和节点偏置采用加入高斯分布的随机数的随机初始化为较小的数值；

(2) 对所有隐层单元$j=1,2\ldots H$，计算$\mathrm{P}(h_{1j}=1|v_1)=\sigma(b_j+\sum_i v_{1i}w_{ij})$，由此抽取$h_{1j}\in\{0,1\}$；

(3) 对所有可见单元$i=1,2\ldots V$，计算$\mathrm{P}(v_{2i}=1|h_1)=\sigma(a_i+\sum_j h_{1j}w_{ij})$，由此抽取$v_{2i}\in\{0,1\}$；

(4) 对所有隐层单元$j=1,2\ldots H$，计算$\mathrm{P}(h_{2j}=1|v_2)=\sigma(b_j+\sum_i v_{2i}w_{ij})$

(5) 更新网络权值和偏置参数：

$$W=W+\varepsilon(\mathrm{P}(h_1=1\mid v_1)v_1^T-\mathrm{P}(h_2=1\mid v_2)v_2^T) \tag{3-13}$$

$$a=a+\varepsilon(v_1-v_2) \tag{3-14}$$

$$b = b + \varepsilon(\mathrm{P}(h_1 = 1 | v_1) - \mathrm{P}(h_2 = 1 | v_2)) \tag{3-15}$$

(6) 若周期小于 T 则返回 (2)，否则结束算法。

Gibbs 采样的示意图如图 3-8 所示。Hinton 指出，虽然这个学习过程没有严格遵循梯度下降的准则，但在实践中采用一步 Gibbs 采样便可得到足够的近似以达到期望效果。

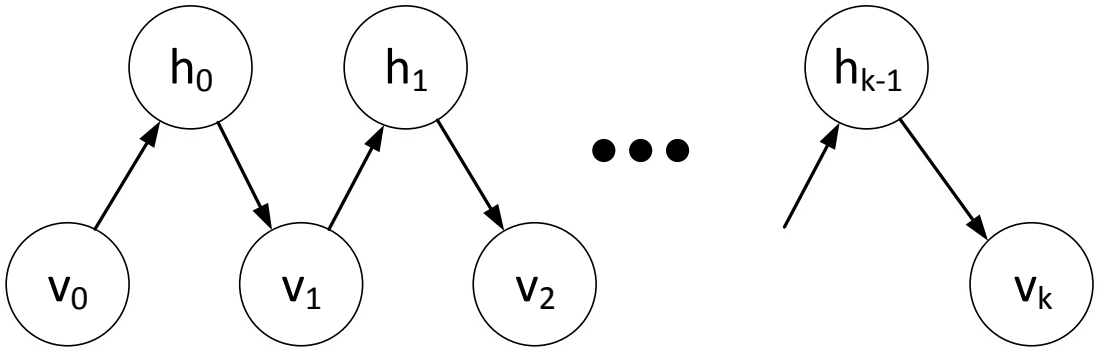

图 3-8 k 步 Gibbs 采样示意图

3.4.2.2 深度信念网络（DBN）模型

经过一次 Gibbs 采样，隐藏层的节点就是对可见层节点的一个重构 (reconstruction)，于是原先在可见层的向量以另一种形式被表示。当多层 RBM 被叠加而形成了一个 DBN 时，这个深度网络便具有很强大的特征学习能力，它将原始的输入层的特征经过每一层 RBM 提取更高一层的特征，并在高维空间对原始数据进行分类。一个两个隐层的 DBN 网络如图 3-9 所示，第一个 RBM 的输出作为下一个 RBM 的输入，这样层层叠加起来的 DBN 是逐层训练的，这要比一次性训练整个深度网络要有效、快速得多。

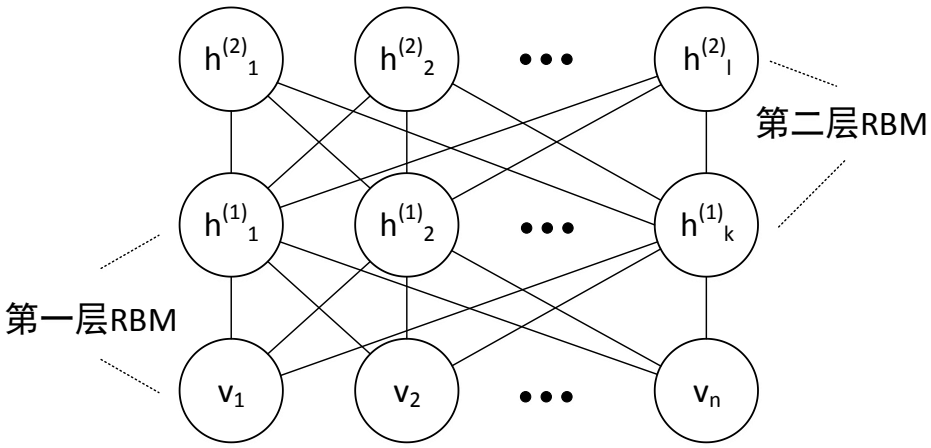

图 3-9 包含两个隐层的 DBN

一个训练好的 DBN 可以用来初始化 DNN 中的权值和偏置参数，这样一来 DNN 已经具备了对训练数据的分类能力，Hinton 建议在此之后使用 BP 算法再对网络权值进行微调以达到最好的效果。这样的生成模型有以下几个好处：

(1) 每一层可以分开训练，逐层叠加；

(2) 可以使用无标注的数据进行训练；

(3) 由于层内节点在给定一个输入时相互独立，这使得公式的推理比较简单，易操作；

(4) 由层层 RBM 叠加成的 DBN 对数据分类具有很好的准确率。

3.4.3 混合 DNN-HMM 分类器

虽然 HMM 在语音处理领域获得了很大的成功，但是一般认为 HMM 还是存在一些缺

陷，比如由于训练准则和训练算法的限制，它对模式识别的能力较差。另外，DNN 具有很强的模式识别和分类能力，并且它不需要对已有的输入样本做任何分布上的假设，所以本课题以 DNN 的输出代替原先 GMM-HMM 模型中 GMM 的输出，从而形成了一个混合 DNN-HMM 分类器。分类器模型示意图如图 3-10 所示。

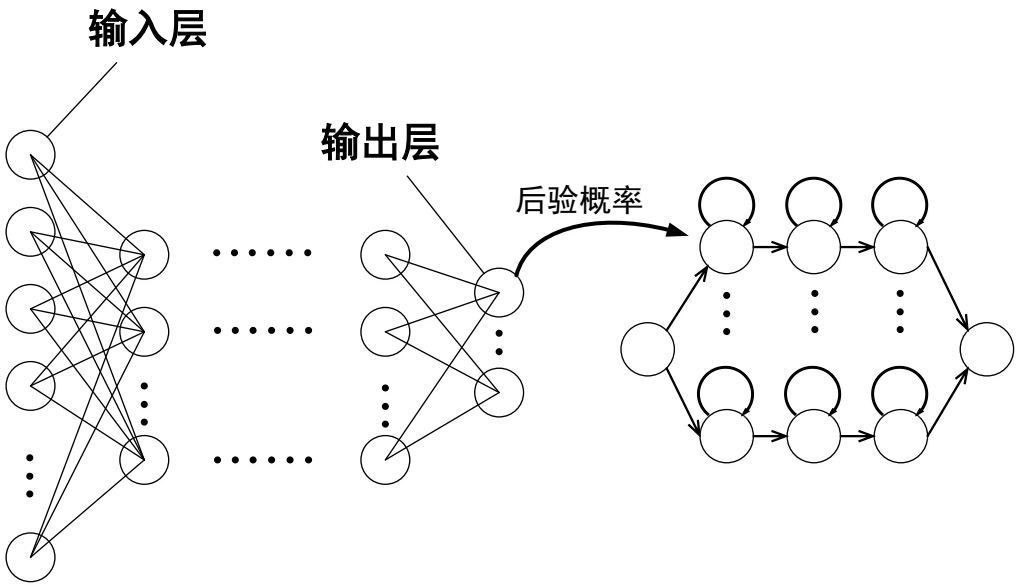

图 3-10 混合 DNN-HMM 分类器示意图

以情感状态分类为例，整个分类器分为两个组成部分：

(1) DNN 模型。DNN 的输入层是之前提取的语音特征参数，如 MFCC、PLP、FBANK 等，在训练 RBM 时，第一层是高斯 RBM，其它的都是伯努利 RBM。DNN 的输出层是对应的情感状态后验概率分布，并且使用 softmax 回归使得所有的节点的输出值的和为 1。

(2) HMM。假设需要进行分类的情感状态的个数为 N，那么便为每一个情感状态建立一个 HMM，为了方便起见，所有的 HMM 都共享一个开始状态和一个结束状态。

在解码阶段，每一帧语音都通过 DNN 进行解码，再将输出层的情感状态后验概率通过相应的 HMM 进行识别，将使得概率$P(Speech|Model)$取得最大值的那个 HMM 所对应的情感状态作为这一个语句的识别结果。

3.5 本章小结

本章主要介绍了基于隐马尔可夫模型的两种分类算法：高斯混合模型和深度神经网络模型。隐马尔可夫模型具有出色的对语音序列建模的能力，在语音处理中，常用的是左右无跳转的连续隐马尔可夫模型。高斯混合分类器是常用的对数据分布进行建模的方法，理论上如果不断地增加高斯模型分量的个数，则可以无限逼近任意地概率分布。但是由于高斯混合模型分类器对语音序列进行分类时，忽略了语音帧之间的前后依赖关系，并且增加高斯分量的个数对计算量也是一个挑战，所以我们尝试使用了深度神经网络模型来进行分类器的搭建。深度神经网络采用了两个训练步骤：

(1) 自低向上的无监督学习，使用无标注的数据训练一个深度信念网络来对网络权值进行初始化，这样的初始化过程得到的网络初值比随机初始化得到的初值更加接近全局最优；

(2) 自顶向下的监督学习，通过带标注的数据对网络权值进行微调，以达到最佳的性能。

其中在无监督学习过程中使用的是受限玻尔兹曼机模型，由于这种模型的特殊内部结构和层次结构，所以可以通过 Gibbs 采样来进行模型的快速训练。受限玻尔兹曼机模型可以层

层叠加而得到深度信念网络模型，这种模型具有学习高维特征的能力，并且对未标注的数据有很好的分类能力。自顶向下的监督学习是使用 BP 算法对深度神经网络进行训 练，深度神经网络的网络初始值是由深度信念网络来初始化，实践证明这比随机初始化网络权值的性能要好很多。最后我们又介绍了结合深度神经网络和隐马尔可夫模型的混合分类器，这种分类器利用了深度神经网络的模式识别能力和隐马尔可夫模型对语音序列的强大建模能力，理论上可以取得比高斯混合模型分类器更好的效果。

第四章 实验和分析

4.1 HTK 工具箱

本课题使用 HTK 工具箱对底层的 HMM 进行建模，运行环境是 Linux 系统。HTK[20] (Hidden Markov Model Tookit) 是由剑桥大学电子工程系语音视觉和机器人研究小组开发，软件是基于 C 编写而成，支持在 Unix/Linux 和 Windows 平台上运行。HTK 工具包可以方便有效地进行基于隐马尔可夫模型的各种实验，它还可以用于字符识别、语音合成、处理 DNA 数据等多个领域。HTK 工具包的软件架构如图 4-1 所示。

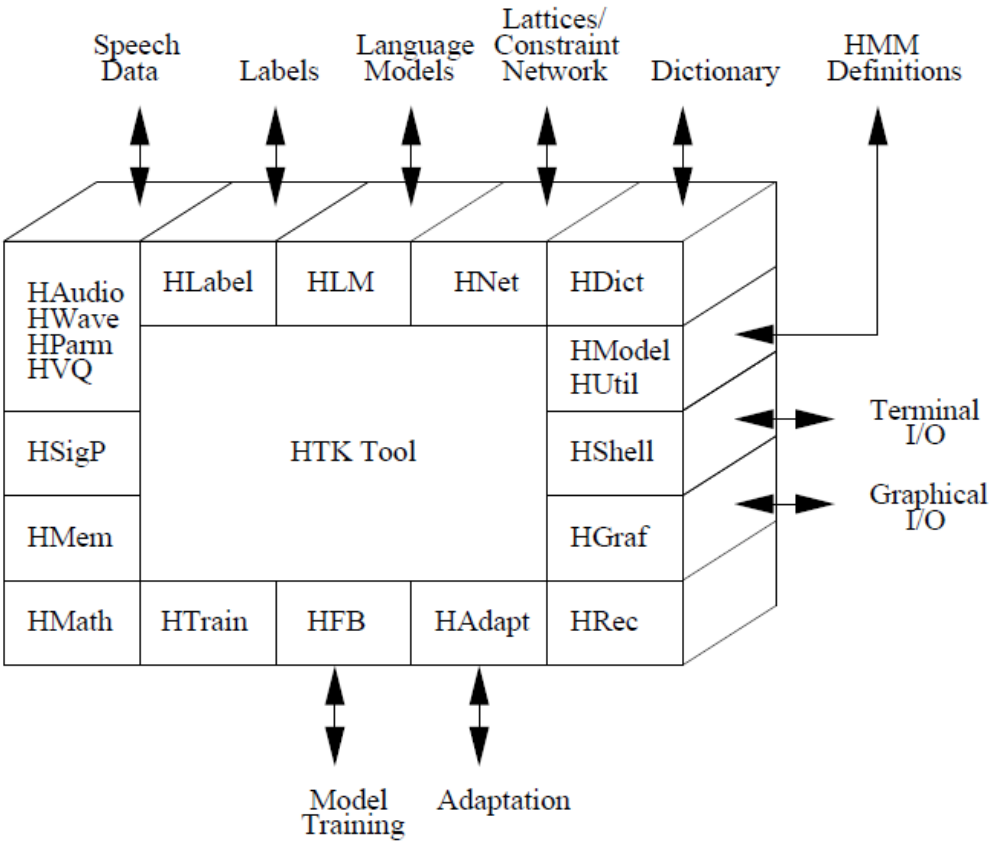

图 4-1 HTK 软件结构图

每个软件模块负责处理不同的数据。比较主要的模块有：

(1) HSHELL：处理用户输入输出以及和操作系统的交互；
(2) HMEM：内存管理模块；
(3) HMATH：数学运算和处理模块；
(4) HSIGP：语音信号的处理模块；
(5) HLABEL：负责处理标注文件；
(6) HLM：负责处理语言模型文件；
(7) HNET：负责处理语言网络；

(8) HDICT：负责处理字典文件；
(9) HMODEL：负责处理隐马尔可夫模型的模型定义；
(10) HWAVE：负责处理.wav 原始音频文件；
(11) HPARM：负责处理将原语音文件转换成参数之后的模块；
(12) HAUDIO：负责处理语音的直接输入；
(13) HREC：包含了识别阶段所用的工具。
使用 HTK 工具包进行语音识别的基本原理如图 4-2 所示。

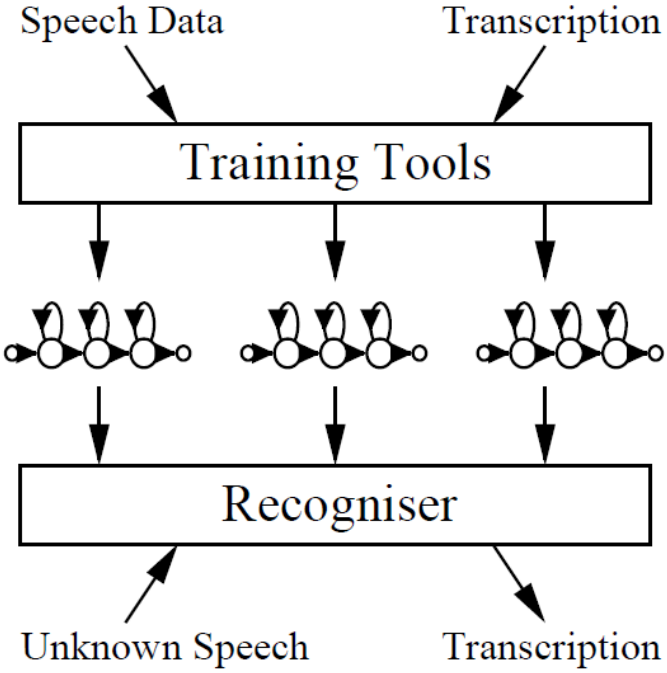

图 4-2 使用 HTK 的基本原理示意图

在训练阶段，使用 HTK 的训练工具对语音数据进行训练，从而构建相应的 HMM 模型。在识别阶段，使用训练好的 HMM 模型对未知的语音数据进行解码识别。其中的各种参数、脚本等都可以通过抄本文件进行控制。

4.2 语音数据的准备

本课题主要研究了对离散情感状态的分类和识别。在实验中，我们总共使用了 9595 句带情感状态的语句，其中使用 7676 句作为训练数据 (15.8% Angry, 15.6% Happy, 17.9% Fear, 19.6% Sad, 15.9% Surprise, 15.4% Neutral)，剩下的 1919 句作为测试数据 (16.7% Angry, 16.7% Happy, 16.6% Fear, 16.7% Sad, 16.7% Surprise, 16.7% Neutral)。数据分布图如图 4-3 所示。

之后我们又做了年龄和方言的分类和识别。年龄识别中使用了 45000 句作为训练数据 (33.3% Young, 33.3% Adult, 33.3% Senior)，4163 句作为测试数据 (44.7% Young, 44.2% Adult, 11.1% Senior)。其中标注为"Young"的年龄段为 15-24 岁，标注为"Adult"的年龄段为 25-54 岁，标注为"Senior"的年龄段为大于 55 岁。年龄的数据分布如图 4-4 所示。

在方言识别中使用了三个标签：重庆、广州和上海。我们使用了总共有 10018 句作为训练数据 (34.9% 重庆, 34.9% 广州, 30.1% 上海)，测试数据根据方言的严重程度又分为 B2 总共 1504 句 (33.8% 重庆, 34.1% 广州, 32.0%, 上海) 和 C 总共 760 句 (31.8% 重庆, 32.7% 广州, 35.4% 上海) 两组，其中 C 组的方言比 B2 组的方言更严重。方言数据的分布图如图 4-5 所示。

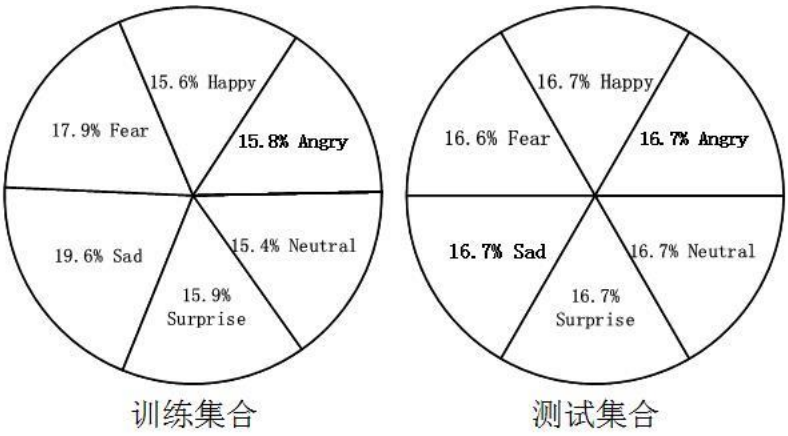

图 4-3 情感数据分布

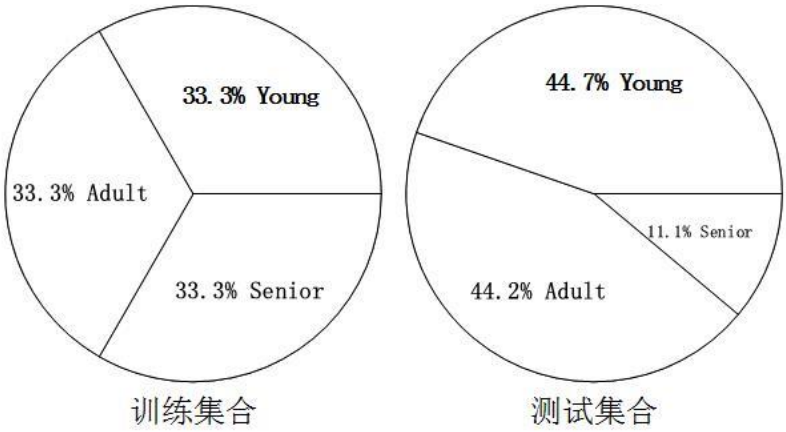

图 4-4 年龄数据分布

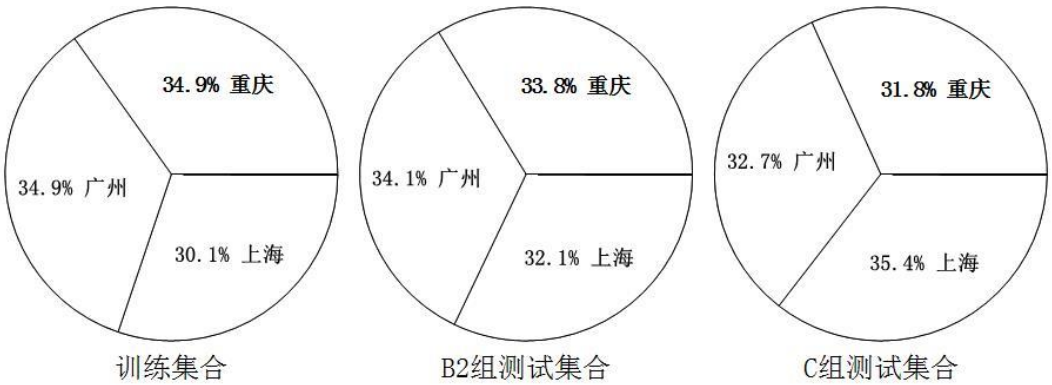

图 4-5 方言数据分布

所有的数据在使用前都去掉了音频首尾的静音部分，只留下可供训练的语音部分。

4.3 特征提取

我们使用 HTK 中的工具对语音数据提取参数，以 MFCC 为例的参数配置文件如表 4-1 所示。此次实验中用到的语音参数有 MFCC、PLP 和 FBANK。其中 MFCC 和 PLP 的基本特征是 13 维，一阶差分和二阶差分动态特征 26 维，总共的特征 MFCC 是 39 维，PLP 也是

39 维；提取 FBANK 参数时设置的滤波器组个数为 24 个，再加上一阶和二阶的动态差分项，FBANK 总共是 72 维。

表 4-1 提取 MFCC 特征时的配置文件参数

参数名称	取值
SOURCEKIND	WAVEFORM
SOURCEFORMAT	WAV
TARGETFORMAT	HTK
TARGETKING	MFCC_0_D_A
TARGETRATE	100000.0
SAVECOMPRESSED	F
SAVEWITHCRC	F
WINDOWSIZE	250000.0
USEHAMMING	T
LPCORDER	12
PREEMCOEF	0.97
NUMCHANS	24
CEPLIFTER	22
NUMCEPS	12
ENORMALISE	T

上述表格中，主要的参数有：

(1) SOURCEKIND 表示源文件的类型，本课题使用的是.wav 文件；

(2) TARGETKIND 表明需要提取的语音特征参数类型，表明需要提取 MFCC 特征并且加上第零个梅尔倒频谱系数，再计算 MFCC 的短时一阶 (Delta) 和二阶 (Acceleration) 差分参数；

(3) TARGETRATE 是指帧移的大小，HTK 中以 100 纳秒为单位，则表格中的数值表示帧移为 10ms；

(4) SAVECOMPRESSED 表示是否将数据压缩存储，F 代表 False；

(5) SAVEWITHCRC 表示是否以 CRC 校验和的方式存储；

(6) WINDOWSIZE 表示窗函数的大小，这里为 25ms；

(7) USEHAMMING 为 T 表示使用汉明窗作为窗函数；

(8) PREEMCOEF 是预加重系数；

(9) NUMCHANS 是滤波器组的个数；

(10) NUMCEPS 是倒频谱系数的个数。

提取 PLP 特征和 FBANK 特征的配置文件和这个类似，却别在于 TARGETKIND 的类型不同。

4.4 分类器的训练

4.4.1 GMM 分类器

HTK 工具箱包含了 GMM 分类器的训练算法，我们使用了最多 512 个高斯分量来逼近训练数据的分布。训练 HMM 时我们采用了三状态和五状态的 HMM。过程如下：

(1) 使用 HCompV 命令得到训练数据的全局方差、原始 flat 文件等，为后面的训练做好准备；

(2) 生成 HMM 定义文件，每个 HMM 定义文件都定义了一个 HMM 模型，其中包括使

用的特征参数的类型以及维度大小、状态转移矩阵、HMM 中有发射概率状态的个数 (在三状态 HMM 中，除去开始和结束状态，所以有发射概率的状态为 1 个，同理，五状态 HMM 中的有效的状态为 3 个)、高斯分量的个数以及每个高斯分量的平均值和协方差，其中协方差是对角矩阵；

(3) 使用 HEREST、HHED 等内置命令对定义好的 HMM 模型来训练。使用 HEREST 调整 HMM 的参数，是用 HHED 增加高斯模型的分量的个数，本实验中采用如下的方法来增加高斯混合模型中的分量：1、2、4、6、8、12、16、32、48、64、128、256、512.。每次增加混合度后对 HMM 的参数重估 4 次。

4.4.2 DNN 分类器

DNN 分类器的训练分为 RBM 的训练和 BP 算法的微调。在数据准备阶段，我们将 15%左右的训练数据划为 CV 集，为后面的数据微调做准备。在 DNN 的输入层，我们采用了语音帧前后扩展 5 帧的方法进行语音帧融合，这样将一帧语音的前后关系考虑了进来，可以使得 DNN 分类器充分学习语音帧之间的依赖关系。DNN 的训练过程如下：

(1) 数据准备。将所有的训练数据随机打乱顺序，并将其中 15%的数据作为 CV 集。之后进行特征的帧扩展，再做全局的归一化；

(2) RBM 训练。将所有的训练数据层层训练 RBM，本实验中训练了 7 层 RBM；

(3) 权值微调。在训练好的 RBM 的最后一层加入输出层，输出层的节点个数根据要分类的状态个数来决定。之后根据标注文件用 BP 算法对网络权值进行微调，直到算法收敛。

在 RBM 训练阶段[31]，学习率设置为 0.02，由于使用了批处理算法，所以批处理的大小设为 256，动量开始设为 0.5，之后提升到 0.9；在微调阶段，批处理大小为 128，学习率设置为 0.8，每次迭代后，都会使用 CV 集对网络进行评估，如果识别率没有很大的上升，则下一轮迭代将学习率减半。

4.5 情感识别结果和讨论

本课题分别使用了 GMM 和 DNN 分类器对情感进行分类。

4.5.1 GMM 分类器识别结果

GMM 实验中，我们训练了两个模型，分别为三状态 HMM 和五状态 HMM，用这两个分类器的结果作为实验的基线结果，以便与之后的深度神经网络分类器的结果作比较。GMM 分类器均使用 MFCC 作为语音特征参数。识别结果的混淆矩阵如表 4-2 和表 4-3 所示，混淆矩阵中每一行之和是 100%。

实验表明，五状态 HMM 的识别率比三状态 HMM 的高 1.4%。其中情感状态 Fear 和 Sad 的识别率较低，这是由于在分类过程中它们被误分类的比例比较大，这也进一步说明 Fear 和 Sad 这两个情感状态有比较相似的谱特征。

表 4-2 三状态 GMM-HMM 分类器对情感的分类结果

	Angry	Happy	Fear	Sad	Surprise	Neutral
Angry	**84.4**	7.5	0.6	1.9	5.0	0.6
Happy	2.5	**82.8**	0.3	0.6	8.4	5.3
Fear	0.6	1.3	**75.2**	18.5	3.3	1.3
Sad	0.6	1.9	17.8	**77.2**	1.3	1.3
Surprise	0.6	4.4	2.2	0.6	**90.6**	1.6
Neutral	0.0	3.8	0.6	0.6	0.9	**94.1**
识别率：**84.1%**						

表 4-3 五状态 GMM-HMM 分类器对情感的分类结果

	Angry	Happy	Fear	Sad	Surprise	Neutral
Angry	**90.3**	2.8	0.3	1.6	4.7	0.3
Happy	2.8	**79.7**	0.3	0.9	12.5	2.2
Fear	0.6	0.9	**80.9**	15.7	1.9	0.0
Sad	0.6	0.9	22.8	**74.7**	0.9	0.0
Surprise	0.6	2.8	1.9	0.3	**94.4**	0.0
Neutral	0.0	5.6	0.0	0.9	0.6	**92.8**
识别率：**85.5%**						

4.5.2 DNN 分类器识别结果

之后我们使用了 DNN 作为分类器。在此次实验中，我们使用了三状态 HMM，而并没有使用五状态 HMM。Le D 和 Provost E M 在他们的工作[9]中表明，在使用混合 DNN-HMM 分类器时，三状态 HMM 的识别率比五状态和七状态 HMM 的识别率都高。并且，他们在实验中主要改变 HMM 的状态数和输入帧数数来讨论对识别率的影响。我们的工作与他们相比主要的不同之处在于，在我们的工作中，我们固定了 HMM 状态数为 3、输入帧数为 11 帧(前后扩展 5 帧)，而改变语音特征类型以及神经网络的大小来讨论对识别率的影响。

当我们使用了 MFCC 为特征参数、隐层数为 1、隐层大小为 1024 个节点的神经网络代替 GMM 后，总体的识别率为 82.3% (混淆矩阵见表 4-4)，这比 GMM 基线的识别率要低一些，但没有低很多。之后我们训练了隐层数为 2-7 层的 DNN，发现识别率大大提高，识别率最高的情况出现在隐层数为 5 的时候，如表 4-5 所示。

表 4-4 ANN 分类器对情感的分类结果

	Angry	Happy	Fear	Sad	Surprise	Neutral
Angry	**88.1**	10.3	0.0	0.9	0.3	0.3
Happy	6.9	**89.1**	0.0	0.3	0.9	2.8
Fear	2.5	2.2	**65.8**	27.0	0.3	2.2
Sad	1.3	1.3	14.1	**81.3**	0.6	1.6
Surprise	5.6	16.0	1.9	0.0	**75.0**	1.6
Neutral	1.9	3.4	0.0	0.3	0.0	**94.4**
识别率：**82.3%**						

表 4-5 5 层 DNN 分类器对情感的分类结果

	Angry	Happy	Fear	Sad	Surprise	Neutral
Angry	**98.4**	0.3	0.3	0.3	0.6	0.0
Happy	1.6	**92.8**	0.0	0.3	3.4	1.9
Fear	0.6	0.0	**79.0**	19.4	0.9	0.0
Sad	0.3	0.3	12.2	**87.2**	0.0	0.0
Surprise	0.3	0.9	0.6	0.3	**97.5**	0.3
Neutral	0.6	0.6	0.0	0.0	0.9	**97.8**
识别率：**92.1%**						

4.5.2.1 改变网络大小对识别结果的影响

对于不同的分类任务，网络的隐层大小会对识别率产生不同的影响。我们又在此基础上改变了隐层的大小，除了之前的隐层大小为 1024 的实验外，我们又分别实验了 512 和 2048

个节点对识别率的影响，综合结果如图 4-3 所示。图中横坐标是隐层的数目，纵坐标是识别率。我们发现当隐层数为 1 时，512 节点数的识别率比 1024 和 2048 节点数的识别率要高，但是随着网络深度的增加，隐层节点数为 512 的 DNN 对情感状态的分类就明显不如 1024 和 2048 两个好了。从隐层节点数为 2048 的识别结果看出，使用两个隐层的 DNN 就已经足够对情感状态进行比较高的识别率，增加隐层数目没有对识别率有很大的提升，而是使识别率趋于稳定。表现最出色的是隐层节点数为 1024 的 DNN 模型，它达到了 92.1%的最高的识别率。隐层大小为 512 的 DNN 的最好识别率是 90.6%，隐层大小为 2048 的 DNN 的最好识别率是 91.6%。

4.5.2.2 改变语音特征对识别结果的影响

之后我们固定了隐层节点数为 1024，并使用 PLP、FBANK 以及 MFCC、PLP、FBANK 的融合来进行实验。图 4-4 展示了不同语音特征对识别率的影响。其中标签”COMBINATION ONE”和”COMBINATION TWO”均表示对特征的融合，只是两个分别使用了不同的方法。下面分别讲解两种特征融合的方法：

(1) COMBINATION ONE：在输入层使用 MFCC、PLP、FBANK 三种特征拼接而成的融合特征，其它参数和条件不变。在此次实验中，得到的用于表示每一个语音帧的特征长度是 150 维 (39+39+72)，从而输入层的大小为 11*150＝1650，一个简单的示意图如图 4-5 所示。

(2) COMBINATION TWO：由于已经对 MFCC、PLP 和 FBANK 做过相关实验，所以可以很容易就得到这三个识别系统中 DNN 的后沿概率。我们对分别使用 MFCC、PLP、FBANK 的识别系统的 DNN 输出求平均值，其它参数和条件不变，并以此平均值作为 HMM 的后验概率。图 4-6 示例了产生情感 Angry 的后沿概率的示意图，其它几个情感的后沿概率的产生过程是一样的。

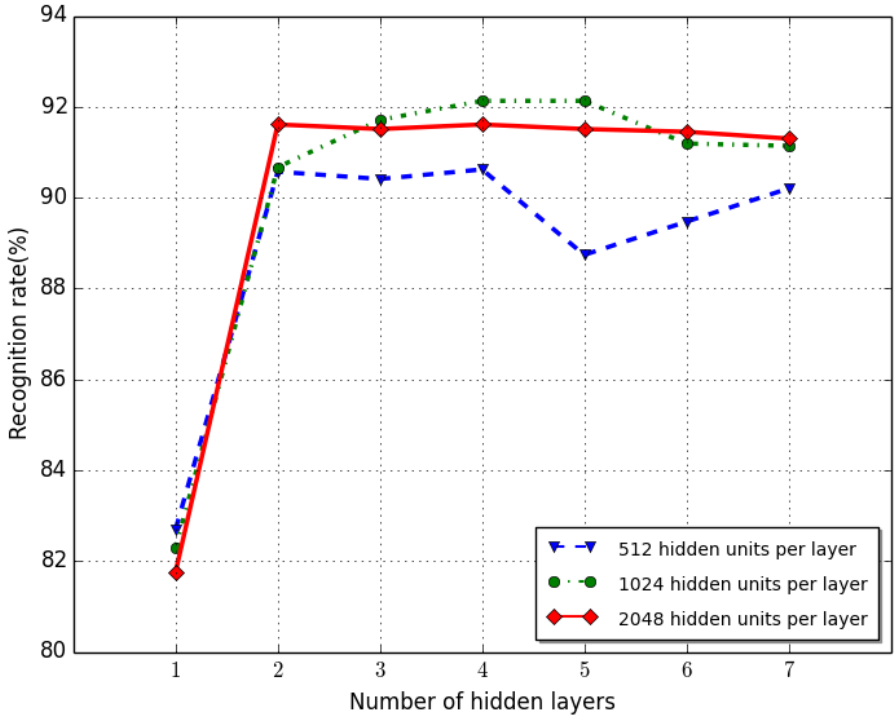

图 4-3 不同隐层大小对情感识别率的影响

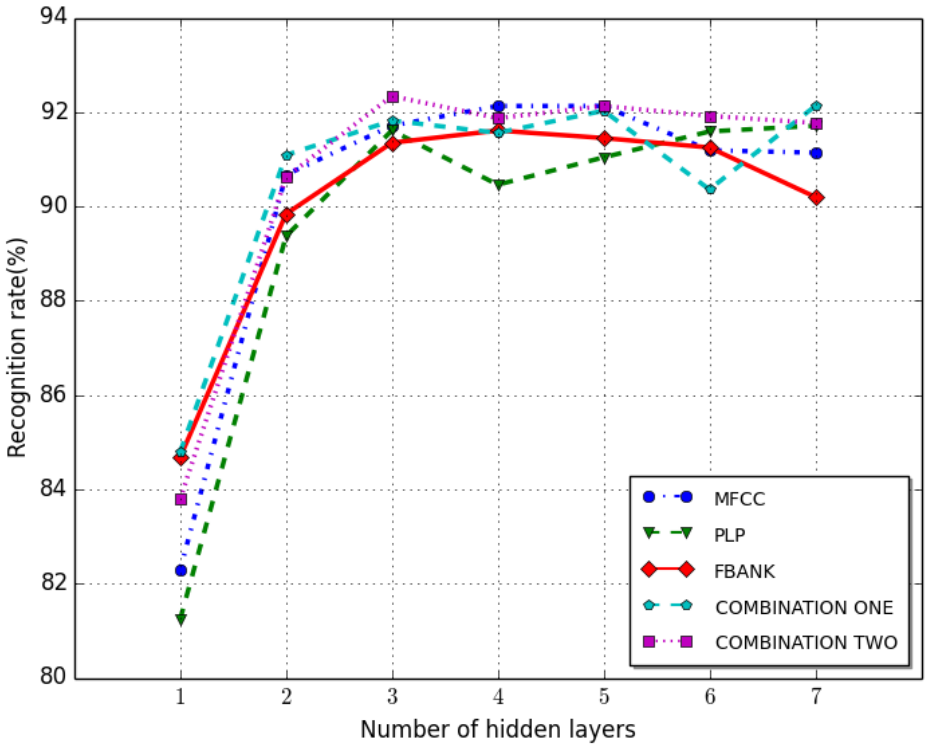

图 4-4 不同语音特征对情感识别率的影响

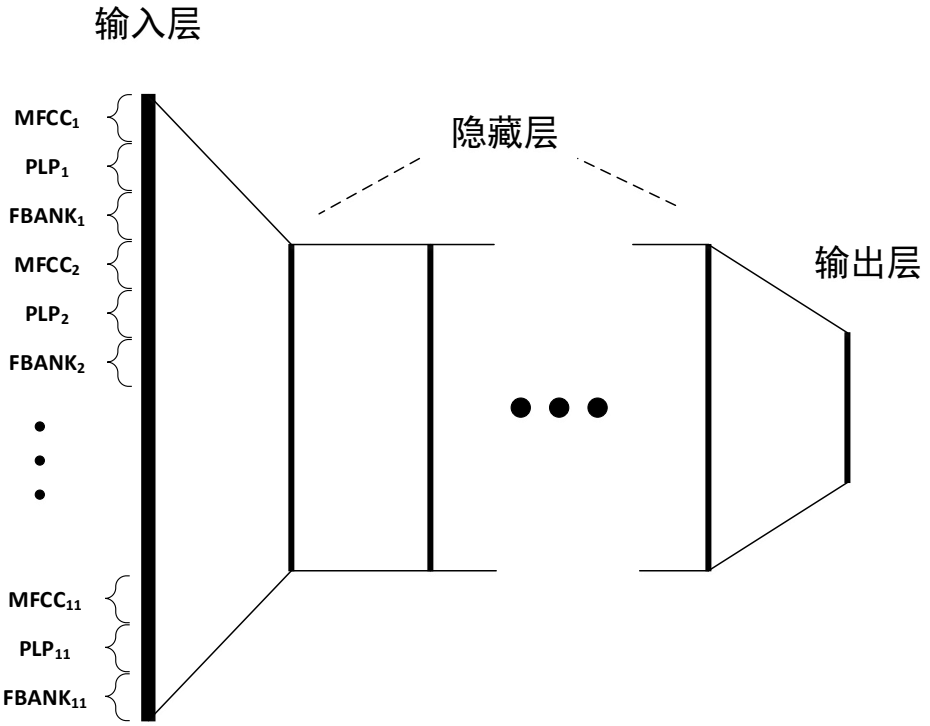

图 4-5 在输入层对三种特征进行融合

实验结果显示，MFCC、PLP、FBANK 都在网络深度增加时表现出了良好的识别能力，而 FBANK 特征的总体表现比较稳定，在隐层数为 1 的时候比其它两个特征有较大的优势。FBANK 的最佳识别率是 91.6%，PLP 的最佳识别率是 91.7%。我们还发现，使用特征融合后的系统识别率比单一特征的识别系统有所改观，其中”COMBINATION ONE”的最佳识别

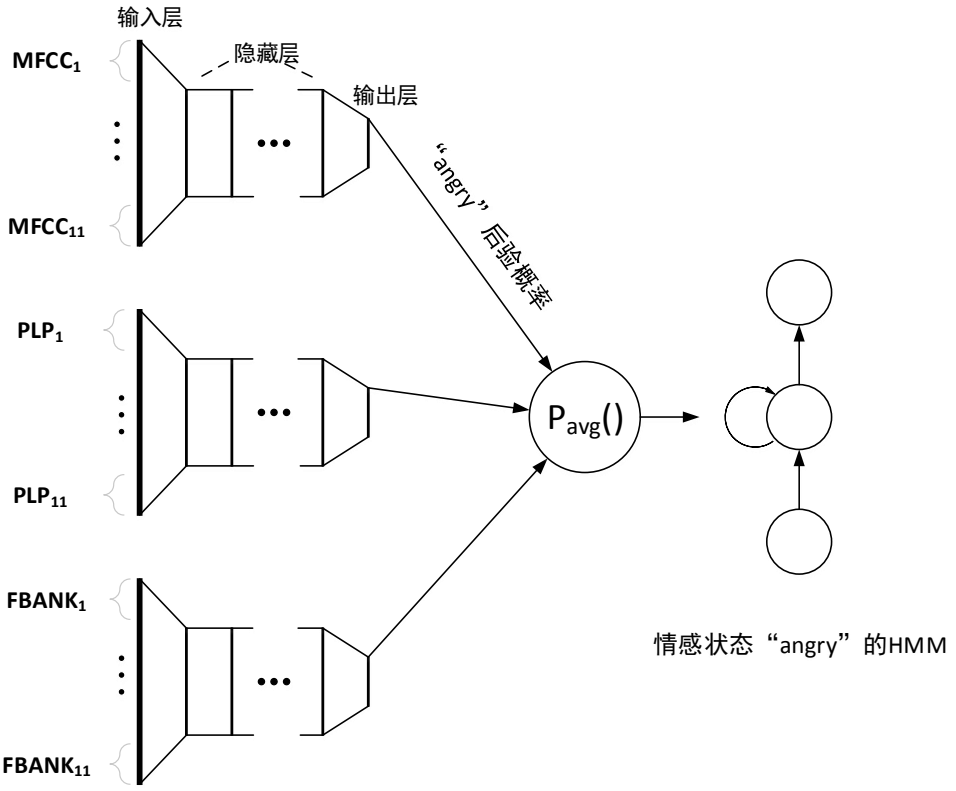

图 4-6 在输出层对三种特征进行融合，以“Angry”为例

率是 92.1%，”COMBINATION TWO”的最佳识别率是 92.3%，这两者都表现出了很好的识别能力，并优于单一特征的识别率。这可以被解释为，使用多种特征对语音进行表示，可以弥补某些特征的在某些方面的不足，以达到提升系统识别准确率的效果[32][33][34]。表 4-6 总结了使用不同分类器和不同特征对情感状态识别率的影响。

表 4-6 不同情感识别系统的结果比较

分类器和语音特征	识别率
DNN，COMBINATION TWO	**92.3%**
DNN，COMBINATION ONE	92.1%
DNN，MFCC	92.1%
DNN，PLP	91.7%
DNN，FBANK	91.6%
ANN，MFCC	82.3%
五状态 GMM-HMM，MFCC	85.5%
三状态 GMM-HMM，MFCC	84.1%

4.6 年龄和方言的识别结果和讨论

对于年龄和方言的识别，我们使用了同样的分类器。使用 GMM 分类器时和之前的参数配置相同；使用 DNN 分类器时，年龄和方言均使用 PLP 特征、输入层的帧长为 11 帧、隐层为 1-7 层。RBM 训练和 BP 算法的参数和情感识别时的参数一致。

4.6.1 年龄识别结果

使用 GMM 分类器对年龄的识别结果如表 4-7 所示，总体的识别率为 55.2%。可以看出，分类器对”Senior”的识别率较低，其中许多被标为”Senior”的标签被分类到了”Adult“中。之

后我们使用了 DNN 分类器又做了分类，结果如图 4-7 所示。

表 4-7 GMM 分类器对年龄的识别结果

	Young	Adult	Senior
Young	**64.9**	31.0	4.1
Adult	32.2	**52.0**	15.9
Senior	7.6	63.0	**29.5**
识别率：**55.2%**			

从图中可以看出，最好的识别率在 60.9%，比使用 GMM 的 55.2%高处了 5.7%。表 4-8 显示了这个最好的识别结果的混淆矩阵。从这个表中可以发现一个现象：虽然总体的识别率比基线 GMM 的要高，但是标签”Senior“的识别率十分低，仅有 2.8%，在识别过程中，大部分的”Senior“标签都被分类到了”Adult“中。DNN 的结果也表现出了一些好处，如标签”Young“的识别率大幅度提高，被误分类到”Senior“标签中的概率也降低了；标签”Adult“的识别率也有所提升，且被误分类到标签”Senior“的数量大大减少。出现这样情况的一个可能性是原始的音频文件中”Adult“和”Senior“具有比较相似的韵律特征和谱特征，从而导致了标签”Senior“被大量地误分类。

4.6.2 方言识别结果

使用 GMM 分类器对方言的识别结果如表 4-9 所示。使用 DNN 分类器的结果如图 4-8 所示。使用 GMM 作为分类器时，B2 组和 C 组的识别率均在 51%左右，总体识别率不高。其中上海方言的识别率较低，广州方言 C 组的识别率较高。DNN 的识别结果中，B2 组的识别率随隐层的变化有一定的规律性：增加到两层隐层时的识别率最高，然后下跌，之后随着隐层的增加识别率有缓慢上升。C 组的识别结果有一定起伏，没有一定的规律性。

表 4-8 6 层 DNN 分类器对年龄的识别结果

	Young	Adult	Senior
Young	**77.7**	22.1	0.2
Adult	39.7	**58.4**	1.85
Senior	7.8	89.4	**2.8**
识别率：**60.9%**			

表 4-9 使用 GMM 分类器对方言的识别结果

	B2 组	C 组
上海	27.6	23.4
广州	64.9	75.1
重庆	60.2	55.0
总识别率	**51.3%**	**50.4%**

从结果来看，B2 组使用 DNN 的识别率甚至不如使用 GMM 的识别率，C 组的识别结果 DNN 的稍微高一些。DNN 对 C 组的识别率要高于 B2 组的识别率，其中广州方言的识别率提高了很多，这和 GMM 的识别结果是一致的。使用 DNN 分类器 B2 和 C 的最好的结果如表 4-10 和表 4-11 所示。

从表中可以看出，对 B2 组和 C 组的分类中，重庆方言和广州方言的识别率较高，上海方言的识别率较低。使用 DNN 没有表现出比 GMM 更多的优势。

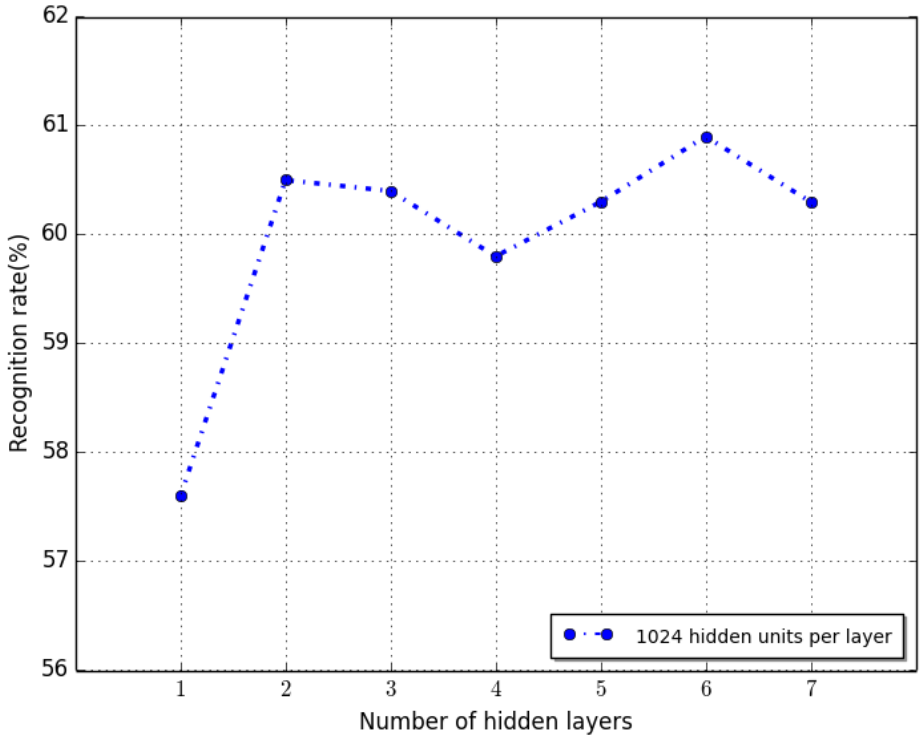

图 4-7 使用 GMM 分类器对年龄的识别结果

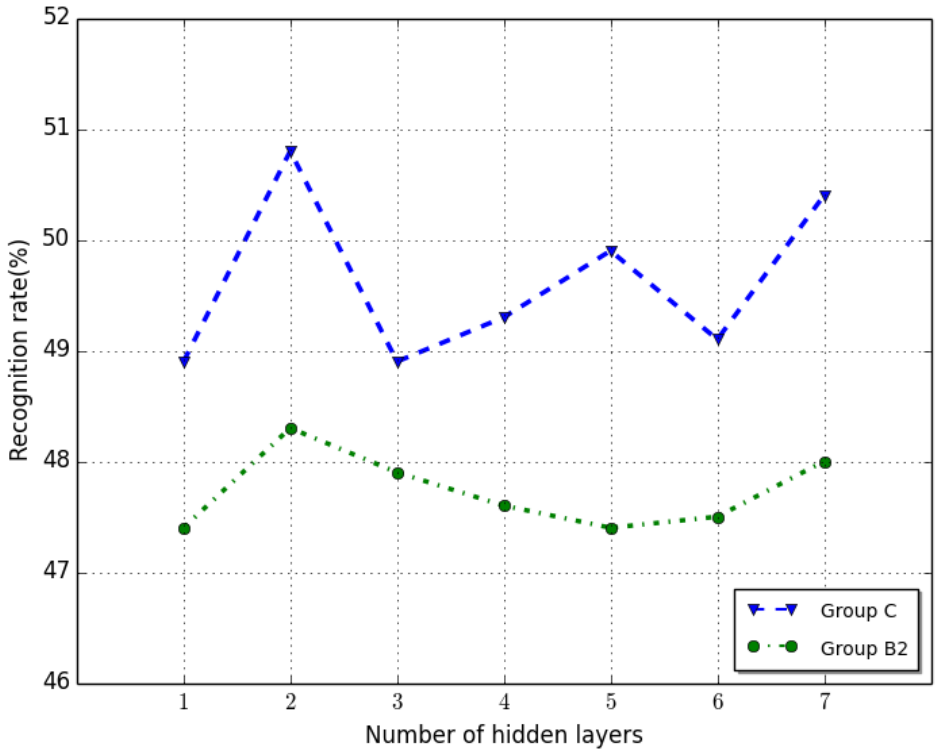

图 4-8 使用 DNN 分类器对方言的识别结果

表 4-10 2 层 DNN 分类器对方言 B2 组的识别结果

	重庆	广州	上海
重庆	**58.0**	41.5	0.6
广州	37.4	**61.4**	1.2
上海	37.3	38.6	**24.1**
	识别率：**48.2%**		

表 4-11 2 层 DNN 分类器对方言 C 组的识别结果

	重庆	广州	上海
重庆	**57.0**	41.7	1.2
广州	26.5	**70.7**	2.8
上海	40.9	32.3	**26.8**
	识别率：**50.8%**		

4.7 本章小结

本章对实验中用到的工具和语音数据做了介绍，并针对情感、年龄和方言做了一系列的实验，以寻找最适合的分类器和分类器参数。HTK 工具包作为处理语音的开源工具，可以很方便地建立 HMM 并进行训练和测试，我们以 HTK 为底层工具，HMM 为基础模型搭建了 GMM 和 DNN 两种不同的分类器。

情感识别的结果显示，使用 MFCC、PLP、FBANK 等参数可以较好地表示情感状态的特点，而使用这三种特征的融合又可以是识别结果进一步提高。使用 DNN 分类器的识别结果比 GMM 分类器有了显著的提高，说明 DNN 在语音情感识别方面可以有很好的表现。

年龄识别中，使用 DNN 分类器的总体识别率要高于使用 GMM 分类器，但在某个单项来看，如标签”Senior“的识别率较低，可能的原因是”Senior“和”Adult“在语音特征上比较相似。如果要提高”Senior“的识别率，可以将三个年龄段的数据更加集中一些，或者重新选择语音特征。

方言识别中，DNN 的总体识别率和 GMM 相当，尤其是上海方言的识别率很低。比较可能的原因是原本的训练数据的筛选，以及对方言的整理不够准确造成的，还有就是测试集数据的筛选不够准确，各个方言的区分度不够等等。要解决这个问题可以考虑对特征做进一步筛选，提取更加适合对方言进行表示的语音特征。

总体来说，我们通过大量试验证明了深度神经网络强大的学习能力和分类能力，它通过在输入层输入多个语音帧，并在层层隐层中提取语音帧之间的依赖关系和原始数据的高维特征，使得对情感标签和年龄标签的识别率得到很大提升。

第五章 总结和展望

本文研究了语音富信息情感、年龄和方言的识别算法，其中深入研究了语音情感的识别问题。语音情感识别是在语音识别中一个热门的话题，从 20 世纪末开始出现了语音情感的数据库，到现在已经有许多优秀的语音情感数据库可以供研究者使用。许多对情感识别的研究都着重于对语音特征的筛选和提取，这包括了韵律学特征如基频、能量、音高等参数，也包括了谱特征参数如 MFCC、PLP 等参数，这说明了在韵律学特征和谱特征中都包含了语音情感的信息。

本课题使用了一种深度神经网络模型对蕴含在语音信号中的富信息进行分类和识别，它的训练过程分为两个步骤：自底向上的无监督学习和自顶向下的有监督学习。已经有许多前人的工作表明，这样的深度神经网络模型具有很好的模式识别能力、分类能力和建模能力，已经在字词识别、语句识别、说话人识别等领域得到了很成功的应用。

我们的所有实验数据来自情感数据库、年龄数据库和方言数据库，并在此基础上做了 GMM-HMM 和 DNN-HMM 的相关分类实验。实验结果表明：

(1) 使用 DNN 作为分类器可以使情感识别和年龄识别结果相比使用 GMM 作为分类器得到很大的提高；

(2) 在情感识别中使用 MFCC 为语音特征参数、DNN 网络大小为 1024 可以达到最佳的识别效果；

(3) 情感识别中三种特征 (MFCC, PLP, FBANK) 的融合可以得到更好的识别效果；

(4) 年龄和方言识别率有待进一步提高。

本课题是对语音富信息情感、年龄和方言识别的初步探索，获得了一些有价值的实验结果，但其中还存在这许多不足之处，日后对此工作的进一步探索可以是：

(1) 本课题仅使用了谱特征，以后的工作可以考虑加入韵律学特征；

(2) 在语音特征的提取过程中，可以对语音特征做筛选、降维，以过滤冗余信息；

(3) 对年龄和方言的识别做进一步研究，有针对性地提取合适的语音特征。其中对方言的数据进行进一步筛选和整理，并设法提高不同方言之间的区分性以提高识别率；

(4) 使用不同的分类器，或混合分类器如 SVM-HMM 等等。

参考文献

[1] 金学成. 基于语音信号的情感识别研究 [D]. 中国科学技术大学, 2007.
[2] 林奕琳, 韦岗, 杨康才. 语音情感识别的研究进展[J]. 电路与系统学报, 2007, 12(1): 90-98.
[3] 陈建厦, 李翠华. 语音情感识别的研究进展[J]. 计算机工程, 2005, 31(13): 35-37.
[4] 赵腊生, 张强, 魏小鹏. 语音情感识别研究进展[J]. 计算机应用研究, 2009, 26(2): 428-432.
[5] 李海峰, 阮华斌, 马琳. 语音情感识别研究进展综述[J].
[6] Stuhlsatz A, Meyer C, Eyben F, et al. Deep neural networks for acoustic emotion recognition: raising the benchmarks[C]//Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, 2011: 5688-5691.
[7] Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using GMMs[C]//INTERSPEECH. 2006.
[8] 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. IEEE, 2013: 3687-3691.
[9] Le D, Provost E M. Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks[C]//Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013: 216-221.
[10]Xia R, Deng J, Schuller B, et al. MODELING GENDER INFORMATION FOR EMOTION RECOGNITION USING DENOISING AUTOENCODER[J]. ICASSP, 2014.
[11]Pohjalainen J, Alku P. Multi-Scale Modulation Filtering in Automatic Detection of Emotions in Telephone Speech[J]. ICASSP, 2014.
[12]Chen Y A, Wang J C, Yang Y H, et al. LINEAR REGRESSION-BASED ADAPTATION OF MUSIC EMOTION RECOGNITION MODELS FOR PERSONALIZATION[J]. ICASSP, 2014.
[13]Amer M R, Siddiquie B, Richey C, et al. EMOTION DETECTION IN SPEECH USING DEEP NETWORKS[J]. ICASSP, 2014.
[14]Deng J, Xia R, Zhang Z, et al. INTRODUCING SHARED-HIDDEN-LAYER AUTOENCODERS FOR TRANSFER LEARNING AND THEIR APPLICATION IN ACOUSTIC EMOTION RECOGNITION[J]. ICASSP, 2014.
[15]Steidl S. Automatic classification of emotion-related user states in spontaneous children's speech[M]. Germany: University of Erlangen-Nuremberg, 2009.
[16]Cowie R, Cornelius R R. Describing the emotional states that are expressed in speech[J]. Speech communication, 2003, 40(1): 5-32.
[17]张智星.Audio Signal Processing and Recognition[Z].http://www.eechina.com/thread-43744-1-1.html.
[18]甄斌, 吴玺宏. 语音识别和说话人识别中各倒谱分量的相对重要性[J]. 北京大学学

报：自然科学版，2001，37(3)：371-378.
[19]邵央，刘丙哲. 基于 MFCC 和加权矢量量化的说话人识别系统[J]. 计算机工程与应用，2002，38(5)：127-128.
[20]Evermann G，Kershaw D，Moore G，et al. The HTK book[M]. Cambridge：Entropic Cambridge Research Laboratory，1997.
[21]王炳锡，屈丹. 彭煊[J]. 实用语音识别基础，2005.
[22]黄岗. 马尔可夫及隐马尔可夫模型的应用[J]. 电子设计工程，2013，21(17)：60-62.
[23]张震南. 人工神经网络技术在语音识别中的应用[J]. 甘肃科技纵横，2008 (4)：21-21.
[24]毛健，赵红东，姚婧婧. 人工神经网络的发展及应用[J]. 电子设计工程，2012，19(24)：62-65.
[25]Hopfield J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the national academy of sciences，1982，79(8)：2554-2558.
[26]吴鸣锐，张钹. 一种用于大规模模式识别问题的神经网络算法[J]. 软件学报，2001，12(6)：851-855.
[27]杨兴，朱大奇，桑庆兵. 专家系统研究现状与展望[J]. 计算机应用研究，2007，24(5)：4-9.
[28]Ackley D H，Hinton G E，Sejnowski T J. A learning algorithm for boltzmann machines*[J]. Cognitive science，1985，9(1)：147-169.
[29]Smolensky P. Information processing in dynamical systems：Foundations of harmony theory[J]. 1986.
[30]Hinton G E，Osindero S，Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation，2006，18(7)：1527-1554.
[31]Hinton G. A practical guide to training restricted Boltzmann machines[J]. Momentum，2010，9(1)：926.
[32]Hermansky H，Ellis D P W，Sharma S. Tandem connectionist feature extraction for conventional HMM systems[C]//Acoustics，Speech，and Signal Processing，2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. IEEE，2000，3：1635-1638.
[33]Bourlard H A，Morgan N. Connectionist speech recognition：a hybrid approach[M]. Springer，1994.
[34]Hinton G，Deng L，Yu D，et al. Deep neural networks for acoustic modeling in speech recognition：The shared views of four research groups[J]. Signal Processing Magazine，IEEE，2012，29(6)：82-97.

致谢

经过将近半年的时间，即将结束对本课题的研究和探索。感谢钱彦旻老师对我的课题的指导和帮助，以及对我的整个课题方向的把握，使得我能够比较顺利并且有针对性地完成任务，并从课题中学到了很多知识。感谢当初俞凯老师向我介绍智能语音实验室，并给予我鼓励和挑战。感谢实验室的陈楠昕，符天凡，刘媛，游永彬这些同学，他们用他们的专业知识为我完成毕业设计的路上提供了很多便捷。语音实验室作为一个刚成立不久的实验室，有着很光明的未来，我也很荣幸能在这个学术氛围良好的实验室完成我的毕业设计。总之很感谢语音实验室的各位老师和同学在做毕业设计期间对我的帮助，也希望智能语音实验能在他们的努力工作下早日建设成世界一流的语音实验室。

四年的本科生涯似乎在弹指一挥间就过去了，其中的喜怒哀乐一时间无法用言语来表达。但我必须感谢周围的同学和朋友们，我们在一起度过了难忘的大学生活；感谢为我们上课的所有的老师们，他们教给我们知识和思想，为我提供学习上和生活上的很珍贵的建议；十分感谢我的父母，在任何时候都能陪伴在我身边，对我的大学生涯给予了极大的支持。感谢你们，见证了我的成长。

学士期间发表论文

[1] Jianwei Niu, Yanmin Qian, Kai Yu. Acoustic Emotion Recognition using Deep Neural Network[J]. International Symposium on Chinese Spoken Language Processing, 2014 已投递

STUDY OF RICH INFORMATION RECOGNITION IN SPEECH

Acoustic speech recognition has been intensively studied in recent years. Besides the recognition of words, phonemes in the speech utterance, there is still much information that we can exploit. In this project, we mainly investigated the recognition of information like the emotion state of a person, the age group of a person, and the accent of a person in a given speech utterance. And due to the time limitation of the project, we mainly focused on the recognition of emotion state from the acoustic speech.

Traditionally there are two kinds of emotion model for describing the emotion state of a person. One is called discrete emotion model. At first, many researchers defined their discrete emotion state relying on their knowledge of emotion recognition. As a result, there were many definitions of discrete emotion states in history made by different researchers. For example, Arnold defined eleven discrete emotion states: anger, aversion, courage, dejection, desire, despair, dear, hate, hope, love, and sadness, while Weiner and Graham just defined two for classification: happiness and sadness. Professor Robert Plutchik proposed an emotion model that has eight basic discrete emotion states: anger-fear, anticipation-surprise, joy-sadness, trust-disgust, while other emotion states were derived from these eight elementary emotion states. This is a popular classification of discrete emotion states in psychology and the most accepted one. In this project, we used discrete emotion states for classification tasks. The other model is dimension model. This model describes emotion in continues space called Valence-Arousal-Power space. The Valence dimension reflects the positive or negative of the emotion, the Arousal dimension the degree of this positive/negative emotion, and the Power dimension describes whether this emotion is active or passive to this person. Most studies of emotion recognition use the discrete emotion states for classification tasks.

Two main focuses, considering emotion classification, are presentation of speech features and type of classifiers. Prosodic features and spectrum features are believed to be very important in emotion recognition. Different emotions have different acoustic features. For example, the emotion states ‘sad’ and ‘fear’ have similar prosodic features, ‘happy’ and ‘excited’ have similar prosodic features. So the emotion states ‘sad’ and ‘happy’ are easy to classify, but emotion states ‘sad’ and ‘fear’ can be easily misclassified to each other due to their similar acoustic features. Spectrum features are also been widely used in tasks of emotion classification, and they are more common when representing the speech utterance.

Usually the speech utterance is segmented into frames of 10ms to 30ms, during which the speech signal is regarded as temporal and stable to be processed. A hamming window is used to be multiplied with the discrete speech signals and then Discrete Fourier Transform (DFT) is performed to get the frame’s discrete spectrum. The speech features are extracted every frame so the original speech utterance is then represented by the sequences of speech features vectors. The most commonly used speech features are MFCCs (Mel frequency cepstrum coefficients). MFCCs are extracted in Mel frequency scale, which is believed to be more suitable to simulate human hearing

system. Many researches were made to proof that MFCCs are better acoustic features when representing the original speech utterance, because the recognition system using MFCCs tended to perform better than recognition systems using other acoustic features. In this project, we used MFCCs as the acoustic features. Besides MFCCs, PLPs (Perceptual Linear Predictive) and FBANKs (filter banks) parameters were also taken into account. PLPs are a kind of parameters that is another form of Linear Predictive coefficient and they are extracted based on the human hearing model. FBANKs are computed directly using the outputs of triangle filter banks arranged in Mel frequency axis. In our experiments, the basic dimension of MFCCs and PLPs were both 13, and the basic dimension of FBANKs was 24 (because we placed 24 filter banks channels). Furthermore, there first and second temporal differences were computed and appended to the feature vectors, thus the final dimension of MFCCs and PLPs were both 39 and FBANKs 72. The dynamic parameters of the features vectors can improve the performance of the recognition system a lot.

Many previous researches reported that the recognition rate of words and phonemes could be improved a lot after deep learning method was employed. In this project, we tried to use deep neural networks to replace the Gaussian mixture models as the classifiers, to see whether our emotion recognition system could benefit from classifiers that using deep leaning. Traditionally acoustic emotion recognition system has been using Gaussian mixture models for classification. However, Gaussian mixture models do not make good use of multiple frames of input data and cannot exploit the high-dimensional dependencies of features efficiently, thus it's hard to improve the recognition rate for achieving a better result. In this project, we introduced a kind of deep neural network as the classifier. This kind of deep neural network is first initialized by pre-training many layers of restricted Boltzmann machines and then fine-tuned using back propagation algorithm. So there are two stages when training the deep neural network. The first stage is unsupervised learning process. In this process, the network is trained using data with no labels. It is very important because in many cases labeled data is not enough when using BP algorithm, so the ability to learn unlabeled data is a major advantage of this network, which is called deep belief network. Deep belief network is stacked up by many layers of restricted Boltzmann machines. Restricted Boltzmann machine is an energy-based stochastic neural network, in which there are no connections in layers. Due to this special structure that the hidden nodes' activation states are conditionally independent of each other when given a particular input vector on the visible layer, it is possible to apply Gibbs sampling to obtain the distribution of the visible vector. There is a fast algorithm to learn restricted Boltzmann machine called Contrastive Divergence which was proposed by Hinton in 2006. This algorithm compute the rough approximation of gradient decent and it works well in practice. Hinton pointed out that one step of Gibbs sampling is enough. Having the help of this fast algorithm, one can train as many layers of restricted Boltzmann machines as wanted. It is also very noticeable that the computation time is linear to the number of hidden layers in the deep belief networks. The second step is discriminatively fine-tuning process. In the first step, the well-trained deep belief network is ready for classification, because it had already learned the complex non-linear relationships between input training data. In this second step, a fine-tune process will be executed to slightly adjust the parameters in the network to better fit the model. Usually BP algorithm is used in this process. This training process of a deep neural network is better, compared with the standard BP algorithm. The standard BP algorithm is easily get stuck in poor local optima especially when there are many hidden layers in the neural network. But to train a deep neural network in the process described above will

avoid local optima and get to global optima, so the performance of the classifier will outperform the original one.

In our project, we did three classification tasks: emotion state, age group and accent classification using both Gaussian mixture model and deep neural network. For emotion recognition, we tested different combinations of input acoustic features and layer sizes. Results showed that all recognition rates using deep neural networks outperformed that using Gaussian mixture models. System with the hidden layer size of 1024 with input features MFCCs gained best accuracy of 92.1%, compared with 84.1% using Gaussian mixture model. We also found that hidden layer size of 512 and 2048 were not as good as 1024. So we fixed hidden layer size at 1024 and changed acoustic features on the input layer. Experiments showed that the overall recognition rate using MFCCs was best. System performance using PLPs and FBANKs was a little lower than those using MFCCs, but they were still much higher than the baselines Gaussian mixture models. The best result using PLPs and FBANKs were 91.7% and 91.6% separately. Combination of three acoustic features helped raised the result to 92.3%, which was best in all of our experiments. In age group classification tasks, deep neural networks improved the baseline results by 5.7%. Deep neural networks did not show much power in accent classification tasks, so in the future it could be a project to investigate accent classification problem.