单位代码：10293　密　级：公开

南京邮电大学

硕 士 学 位 论 文

论文题目：基于多任务神经网络的多维语音识别技术研究

学　　号	1017010525
姓　　名	冯天艺
导　　师	杨震　教授
学 科 专 业	信号与信息处理
研 究 方 向	语音信号处理与语音通信
申请学位类别	工学硕士
论文提交日期	2020.4

Research on Multi-dimensional Speech Recognition Technology Based on Multi-task Neural Network

Thesis Submitted to Nanjing University of Posts and Telecommunications for the Degree of

Master of Engineering

By

Feng Tianyi

Supervisor: Prof. Yang Zhen

April 2020

摘要

二十一世纪以来，信息技术的发展日新月异，在人工智能的浪潮下，实现简单、快捷、流畅的人机交互成为人们追求的目标。通过语音实现交互一直是人机交互领域重要的一部分，而语音识别技术正是人机语音交互的关键技术。近年来，研究者们在语音识别领域做了许多工作，取得了颇为丰硕的成果。真实环境中的语音信号是复杂的混合信号，其中既包含了丰富的语义信息，也包含了许多说话人相关信息（如身份、情感等）和环境信息，这也是我们人类能够顺畅沟通的前提。然而，目前绝大多数的语音识别研究主要集中在针对某单一内容或信息的识别，几乎没有研究能够像人一样同时识别语音信号中包含的多维信息。这样的单维语音识别模型忽略了人脑对多维语音信息的处理能力，摒弃了语音混合信号中多维信息之间的相关性，不利于机器理解语音的真正含义，也不符合智能化人机交互的要求。因此，为了使语音识别技术能够更加拟人化、智能化，本团队提出了对语音信号中的多维信息同时进行识别的课题，充分利用语音信号中丰富的多维信息，挖掘不同语音信息之间的相关性，对多项语音识别任务进行同时分类。本文在本团队前期研究的基础上，从分类模型构建和特征提取两个方面入手，研究说话人性别、情感、身份三类语音信息的同时识别。本文的主要工作和创新点如下：

（1）本文将多任务学习（Multi-task learning, MTL）机制与循环神经网络（Recurrent Neural Network, RNN）结构相结合，充分利用语音信号中丰富的多维信息以及不同识别任务间的相关信息，构建了一个可以同时识别说话人性别、情感、身份的多维语音识别模型。模型采用梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficient, MFCC）特征作为语音识别特征参数，选取带有属性依赖层的多任务神经网络结构，通过 RNN 共享层共享网络参数以学习各识别任务间的共有特征，通过全连接属性依赖层学习各识别任务自身的独有特征，利用 MTL 的机制调整模型总损失函数中各识别任务损失函数的权重，来针对语音数据库的特点进行性能优化，最终同时输出三种识别任务的识别结果。经过多项对比实验的验证，结果表明本文提出的基于 MTL 和 RNN 的多维语音识别模型在两种语音库上的平均识别率分别比单维识别高出 3.01% 和 5.09%，三项识别任务均有一定的识别率提升，对于语种因素和说话人的个性因素有较好的鲁棒性，且具有一定的抗噪性能。不但展示多维任务识别的可行，同时也证明不同任务之间具有明显相关性，多维识别也是提高单维任务性能的重要方法。

（2）由于基于 MTL 和 RNN 的多维语音识别模型采用的语音特征是常用的 MFCC 特征，其在特征提取时的各种滤波和变换操作去除了部分语音信息，而多维语音识别要求尽可能多的利用到语音信号中的多维信息。因此，本文将卷积神经网络（Convolutional Neural Network,

CNN）结构与特征融合（Feature fusion）方法相结合，对特征提取部分进行改进，构建了一个基于 CNN 和特征融合的多维语音识别模型。将语音信号语谱图经过 CNN 提取的特征和人工提取的 MFCC 特征进行融合，充分利用了语音信号中的多维度信息，使两种特征进行互补，最后使用融合特征输入到多任务循环神经网络分类器中完成说话人身份、性别、情感三项任务的识别。经过实验验证，结果表明本文提出的基于 CNN 和特征融合的多维语音识别模型在两种语音库上的平均识别率分别比单维识别高出 3.59%和 6.01%，比 MTL-RNN 模型高出 0.85%和 0.99%，三项识别任务均有识别率提升，且具有更好的抗噪性能，证明了融合特征在多维语音识别上的有效性。

关键词： 多维语音识别，多任务学习，循环神经网络，卷积神经网络，特征融合

Abstract

Since the 21st century, information technology has been greatly developed. Under the tide of artificial intelligence, the realization of simple, fast and smooth human–machine interaction has become the goal of researchers. Speech communication is an important part of human-machine interaction, and speech recognition is the key technology of human-machine voice interaction. In recent years, researchers have made a lot of efforts in speech recognition and obtained rich achievements. Speech signals in real environments are complex mixed signals which contain rich semantic information, speaker-related information (such as identity, emotion, etc.) and environmental information. It is the premise that human beings can communicate smoothly. However, most current studies in speech recognition focus on a single task such as speech content recognition, and it is difficult to recognize multi-dimensional information contained in speech signals simultaneously like human beings. However, such single-dimensional speech recognition models ignore the ability of the human brain to process multi-dimensional speech information and abandon the correlation among multi-dimensional information in speech signals, which is not good for machines to understand the true meaning of speech and cannot meet the requirements of intelligent human-computer interaction. Therefore, in order to make speech recognition technology more anthropomorphic and intelligent, our team proposes simultaneous recognition of multi-dimensional information in speech signals which realizes multiple speech recognition tasks by making full use of the rich multi-dimensional information in speech signals and the correlation among different recognition tasks. Based on the previous research of our team, this thesis studies the simultaneous recognition of speaker's gender, emotion and identity from the aspects of classification model construction and feature extraction. The main work and contributions of this thesis are summarized as follows:

(1) By combining the multi-task learning (MTL) mechanism with the recurrent neural networks (RNN) structure, this thesis makes full use of the rich multi-dimensional information in speech signals to build a multi-dimensional speech recognition model that can simultaneously identify the speaker’s gender, emotion and identity. The model adopts Mel-Frequency Cepstral Coefficient (MFCC) features as the recognition parameters and builds a multi-task neural network structure with attribute dependent layers. It learns common features among different recognition tasks and unique features of each recognition task through the RNN sharing layer and the full connection

attribute dependency layers, respectively. In addition, it leverages the MTL mechanism to adjust the weight of each recognition task's loss function in the total loss function of the model for optimizing the performance according to the characteristics of the speech database, and finally outputs the recognition results of three recognition tasks simultaneously. The experimental results on two speech databases show that the proposed multi-dimensional speech recognition model based on MTL and RNN leads to 3.01% and 5.09% higher average recognition rates as compared to the single-dimensional speech recognition model, respectively, and obtains significant improvement in all the three recognition tasks. Moreover, it is shown that the proposed model is robust to language and speaker's personality factors, and has certain anti-noise performance. In addition, it not only shows the feasibility of multi-dimensional task recognition, but also proves that different tasks have obvious correlation so that multi-dimensional recognition is an important method to improve the performance of single-dimensional tasks.

(2) Since the proposed multi-dimensional speech recognition model based on MTL and RNN adopts MFCC features, it would remove part of speech information in various filtering and transformation operations during feature extraction. However, multi-dimensional speech recognition requires that the usage of multi-dimensional information in speech signal should be as much as possible. Therefore, this thesis further proposes to combine the structure of convolutional neural network (CNN) with the feature fusion method for improving the feature extraction, and construct a multi-dimensional speech recognition model based on CNN and feature fusion. Specifically, the features extracted by CNN and MFCC are fused to make full use of the multi-dimensional information in the speech signal to form complementary features. Then, by using the fusion feature as the input of the multi-task recurrent neural network classifier, the identification of the speaker's identity, gender and emotion is realized. The experimental results on two speech databases show that the average recognition rate of the proposed multi-dimensional speech recognition model based on CNN and feature fusion is 3.59% and 6.01% higher than that of the single-dimensional speech recognition model, respectively, and 0.85% and 0.99% higher than the MTL-RNN model, respectively. Moreover, the results show that the model can improve the recognition rate in all three recognition tasks and has better anti-noise performance, which proves the effectiveness of fusion features in multi-dimensional speech recognition.

Key words: Multi-dimensional speech recognition**,** Multi-task Learning**,** Recurrent Neural Networks, Convolutional Neural Networks, Feature fusion

目录

专用术语注释表

缩略词说明：

ANN	Artificial Neural Network	人工神经网络
BP	Back Propagation	反向传播
CNN	Convolution Neural Network	卷积神经网络
DBN	Deep Belief Network	深度置信网络
DCT	Discrete Cosine Transform	离散余弦变换
DFT	Discrete Fourier Transform	离散傅里叶变换
DNA	Deoxyribonucleic Acid	脱氧核糖核酸
DNN	Deep Neural Network	深度神经网络
DTW	Dynamic Time Warp	动态时间规整
GMM	Gaussian Mixture Model	高斯混合模型
GRU	Gate Recurrent Unit	门循环单元
HMM	Hidden Markov Model	隐马尔科夫模型
i-vector	Identity Vector	身份矢量
LPC	Linear Predictive Coding	线性预测编码
LPCC	Linear Predictive Cepstrum Coefficients	线性预测倒谱系数
LSTM	Long Short-term Memory	长短期记忆
MFCC	Mel-Frequency Cepstral Coefficients	Mel 频率倒谱系数
MIML	Multi-Instance Multi-Label learning	多示例多标记学习
MTL	Multi-Task Learning	多任务学习
ProgNets	Progressive neural networks	渐进式神经网络
SIANN	Shift-Invariant Artificial Neural Networks	移不变人工神经网络
SVM	Support Vector Machine	支持向量机
UBM	Universal Background Model	通用背景模型
VAD	Voice Activity Detection	语音活动性检测
VQ	Vector Quantization	矢量量化

图表说明

图说明：

表说明：

第一章 绪论

1.1 研究背景与意义

二十一世纪以来，人类社会步入信息时代，科学技术的发展日新月异，突破性的成果层出不穷，同时人们的日常生活质量也在不断改善。如今，在人工智能的浪潮下，实现人与机器间流畅高效的交互成为人们追求的目标。众所周知，语音是人与人之间最自然、快捷、有效的交互方式之一，通过语音实现交互一直是人机交互研究领域中重要的组成部分。完善的人机交互要求计算机认证用户身份、理解用户的意图和情绪，从而给予不同的反馈和支持[4]。这一目标的实现首先要求计算机能够准确识别出语音中所包含的各类语义信息和说话人信息，例如话语内容、说话人身份、性别、年龄、情感等，甚至背景声音环境信息。因此，语音识别技术是实现人机语音交互的关键技术之一。人耳接收到的语音信号实际上是一种混合信号，其中主要包含三大类信息，分别是语音内容信息即语义、说话人身份等特征信息、说话人所处的背景环境声音信息。研究者们把针对语音信号中各类信息的识别研究归纳为广义的语音识别技术，包括语音文本信息识别、说话人身份识别、说话人情感识别、语音搜索、语种识别等[1]。目前，语音识别技术已经在各个领域内得到了广泛的应用。

在语音识别领域，已有不少优秀的研究者做了许多工作，也取得了丰硕的成果，但就目前的语音识别研究而言，其研究成果往往是针对某单一信息或内容的识别。然而从宏观上来看，所谓“人工智能”的最终目的必然是“机器拟人化”，在语音识别领域反映为计算机像人一样接收、分析、识别语音信号。众所周知的是，人耳在听到语音信号时（此处指包含语音本身和环境声音等在内的混合声音，下同），大脑会同时产生对多种信息的判断，例如听到的话语的内容、说话人的性别、年龄、情感、身份、说话人所处环境等。在极短的时间内，人脑可以对听到的语音同时进行多维度信息的识别分析。显而易见的是，现有的单一信息类型的语音识别研究忽略了人脑对多维信息的处理能力，也摒弃了语音混合信号中多维信息之间的相关性，是难以满足“拟人化”要求的。各类语音信息的单独识别既不利于机器理解语音的真正含义（例如说话人的不同情感，对应同样语义的话时可能真正的含义不同），也不利于语音识别鲁棒性的提高，更不利于人机语音交互技术的未来发展。因此，为了实现语音识别的智能化和拟人化，本团队创新地提出了多维语音信息同时识别的课题，并取得了初步的成果[5-7]。多维语音识别的基本理念是充分利用不同语音信息之间的相关性，对语音信号中的多种语音信息进行同时识别。由于其所包含的识别任务过于宽泛，且目前缺乏成熟的文献资料

和理论推导，因此本文在本团队已有研究的基础上，基于多任务学习理论，首先选取说话人性别、情感、身份三种语音信息进行联合识别研究，未来再推广到更多的语音信息识别任务中。

1.2 语音识别研究现状

目前，语音识别技术在通信、医疗、服务业、教育等领域已经得到了广泛的引用。其中说话人身份识别和语音内容识别应用最为广泛，可用于语音输入、同声传译、智能家居、地图导航、身份认证等等。随着生活水平的提高，消费者们对于人机交互的要求也在不断升高，机器需要全方面了解用户的情况，理解用户的意图、情绪，才能够提供更加人性化的服务。这就对语音识别技术提出了新的要求，传统语音识别的局限性也逐渐显露出来。因此，将传统的单维语音识别研究扩展到多维语音识别研究，是人机语音交互未来的研究和发展方向。

到目前为止，传统的单维语音识别技术已经有了很长的研究历史，也产生了许多成熟的技术和成果应用。要搭建一个可以同时识别多项语句信息和说话人信息的多维语音识别系统，首先要对各单一识别任务的技术有深入的了解，通过分析不同识别任务间的相关性和差异性，找出缺陷和不足，才能搭建出更加合理的多维语音识别系统。这里介绍本文研究中包含的三个常见的语音单维识别技术：说话人性别识别、情感识别、身份识别。

（1）说话人身份识别技术

说话人身份识别又名声纹（Voiceprint）识别，由于每个人的发声器官具有差异性，造成了每个说话人发出的语音具有唯一性，类似每个人拥有独特的指纹一样，因此叫做声纹。声纹和指纹一样，都属于人类自身的生物特征，因此说话人身份识别技术也是一种生物特征识别技术[8]，又称生物认证技术。在科技飞速发展的现代社会，需要用到身份认证技术的场景越来越多，传统的身份认证方式例如姓名、身份证、证件号码、密码等方式均存在容易丢失、复制、伪造、冒充的安全性问题，而生物认证技术以其可靠、获取难度低、独特、难以伪造、不可复制的特点逐渐成为身份认证技术的主流，是当今社会必不可少的主流技术之一。而在生物认证技术中，说话人身份识别技术又因为其便捷性和动态可变性而具有独特的优越性。相对于指纹、DNA、面部识别等特征，语音更加容易采集，仅仅需要一个麦克风或者通过远程的手机，采集被认证人的一句话即可，更加重要的是，系统可以通过文本无关识别技术，要求说话人讲动态变化的特定的话而使得假冒者难以事先通过录音的方式伪造。因此，说话人身份识别技术一直是业界研究的热点。

研究者们按照识别目的的不同把说话人身份识别分为说话人确认（Speaker Verification）

和说话人辨认（Speaker Identification）两种类别。说话人确认较为简单，属于 0/1 问题，只需针对一个特定的说话人训练一个模型，再将测试所用的语音输入模型，判决得出是否为此说话人，其结果只能是‘是’或者‘否’。说话人辨认则更加困难，属于 1/*N* 问题，需要从 *N* 个说话人中辨认出某个测试语音的说话人身份。此外，还可以按照语音的内容是否可变分为文本相关（Text-Dependent）和文本无关（Text-Independent）两种类别。文本相关指的是识别模型对输入的训练和测试语音数据的文本内容有要求，说话人需按照设定的内容发声，这种方式局限性较大，扩展性差，但识别性能较高，适合说话人配合的场景[9]。而文本无关的说话人身份识别则对语音数据的内容没有要求，适用范围更广，同时模型的建立更加困难和复杂。

最早的说话人身份识别研究可以追溯到二十世纪三十年代，人们意识到通过声音确定身份的技术在很多场景下都有很好的应用。1945 年，大名鼎鼎的贝尔实验室的 Kesta 首次提出“声纹”的概念[10]，通过语谱图的观察和匹配来试图分辨不同说话人[11]。从那以后，研究者们开始使用语音中的特征参数来进行说话人身份识别，并加入了统计分析方法，取得了很大的进展，人们开始向着机器的自动说话人识别的方向不断前进，也掀起了业界对说话人身份识别研究的高潮[12]。那时研究者们对说话人身份识别的研究主要集中在语音特征参数的提取上。1963 年，Bogert 等人基于倒谱分析进行说话人身份识别，并取得了很好的效果。1974 年，Atal 将线性预测倒谱系数（Linear Predictive Cepstrum Coefficients, LPCC）运用在说话人身份识别上，大大提高了识别性能[13]。在这之后，大量的特征参数被发掘使用，例如 MFCC、共振峰、高阶统计特性等。后来的研究逐渐从语音特征参数转到模式匹配的研究上[1]。动态时间规整（Dynamic Time Warping, DTW）[14]、矢量量化（Vector Quantization, VQ）[15]、隐马尔科夫模型（Hidden Markov Model, HMM）等技术成为当时研究的主流，至今甚至仍有使用。二十世纪九十年代以后，高斯混合模型（Gaussian Mixture Model, GMM）、支持向量机（Support Vector Machine, SVM）[16]、人工神经网络（Artificial Neural Networks，ANN）[17]等技术也逐渐被用于说话人身份识别中，不断提升识别性能。2011 年，Dehak 等人提出了身份矢量（i-vector）模型[18],大大提高了识别性能。近来，随着人工智能和机器学习的发展，基于深度神经网络的方法也被应用到说话人身份识别中，并取得了不错的效果[19]。

目前，随着信息时代的来临，人类社会对身份认证、信息安全愈发重视，说话人身份识别得到了广泛的应用，在金融、电信、教育、安保、军事等各领域都有其应用场景。可以预计，说话人身份识别技术具有广阔的发展前景，在未来的很长一段时间中都将会是研究的热点。现阶段的说话人身份识别研究已经取得了一定的成果，但仍存在一些问题亟待解决。例如，在说话人声音改变的情况下（如感冒、紧张、变声软件等）如何准确识别出说话人的身

份；在语音长度较短时如何提高识别准确率；在环境噪声较大的公共场合如何保证识别性能等。因此，说话人身份识别还需要进一步的研究和发展。

（2）说话人性别识别技术

说话人性别识别技术指的是计算机通过分析语音特征来判断说话人性别的技术，是语音识别领域中较为基础的一部分。与其他语音识别技术相比，性别识别只有两种识别结果，难度相对较低，算法实现更为简单，且有研究表明，在说话人身份识别领域中，基于男女分类前提下的说话人身份识别系统的识别率和运行速度，都要显著优于未进行性别分类的说话人身份识别系统，因此最初的说话人性别识别是说话人身份识别的子领域之一[20]。

直到 1988 年，Childers 等人发表在 ICASSP 上的论文首次将说话人性别识别从说话人身份识别中分离出来进行专门的研究，成为一个独立的研究领域[19]。最初，说话人性别识别的研究主要以特征参数的筛选和提取为主，其中，基音频率和共振峰是最主要的识别特征。通常男性正常讲话时的基音频率大致在六十赫兹到两百赫兹之间，而女性正常讲话时的基音频率往往比男性更高，最高可以达到四百五十赫兹左右，因此通过基音频率可以进行大部分情况下的判决。同样，女性语音的共振峰频率也明显比男性语音要高。后来，研究者们逐渐加入了韵律特征、LPCC、MFCC 等特征，也逐渐从音节、孤立词的判别转为短时语音、连续语音序列的识别[21]。在分类器的选择上，从最初的隐马尔科夫模型到后来的高斯混合模型、支持向量机，再到近年来深度学习方向的各类神经网络模型，分类算法的不断发展也使得说话人性别识别的准确率和鲁棒性不断提升。

目前，说话人性别识别技术已经较为成熟，在纯净无噪声的语音环境下的性别识别准确率无限接近于百分之百。同样，性别识别研究也存在着一些难点，例如童声的性别识别、低信噪比条件下的性别识别等。如今的性别识别研究正在朝着实用化的方向发展，一方面专注于各类复杂背景噪声环境下的鲁棒性性别识别，另一方面往往将性别识别作为其他语音信息识别技术研究的基础或一部分，利用性别相关性，提升其他语音识别任务的性能[22,23]。

（3）语音情感识别技术

前文提到，完善的人机交互要求计算机理解用户的情绪和意图，从而给予不同的反馈和支持[4]，因为情感反映了说话人的心情和态度，也与真实语义密切相关。而实现这一目的首先要求计算机能够准确的识别出说话人的情感，因此语音情感识别在人机交互中具有重要的意义。最早的真正意义上的语音情感识别相关研究出现在 20 世纪 80 年代中期，研究者们试图让计算机分析说话人的语调和语气，判别不同种类的情感状态，由此开创了使用声学统计特征进行情感分类的先河[24]。早期的语音情感识别往往通过提取低级特征来训练分类模型，例如支持向量机、隐马尔科夫模型等[25,26]；也有一些研究者针对每一种情感训练一个高斯-隐

马尔科夫混合模型，并在测试阶段将话语分类为可能性最大的情感标签[27]。这些研究中，语音特征参数的提取是否合理往往直接影响模型识别性能的好坏。语音情感识别常用的特征有能量、基音频率、过零率、共振峰、LPCC、MFCC 等，这些特征各自从不同侧面描述了语音中包含的情感信息，实际研究中往往采用这些特征的融合特征进行情感分类。另外，关于情感的分类也有两大主流方式，分别是离散情感模型和维度情感模型[28]。离散情感模型把情感分为特定的几类情感，常见的有快乐（happy）、悲伤（sad）、生气（anger）、吃惊（surprise）、平静（neural）、害怕（fear）等类别，也有更细致的划分方式，例如快乐可以分为幸福（happy）、高兴（pleased）、愉悦（cheerful）等，越细分识别难度越大；维度情感模型认为情感没有严格的界限，应该用连续的空间维度来表示，每一个空间维度都代表着情感的一个心理学属性，维度的数值大小反映了情感在相应维度上表现出的强度[24]。两种情感模型各有优劣，前者有较大局限性，但结果较为简单明确；后者描述能力更强，但模型复杂度较高，且没有统一的定义方式。

近年来，随着计算机性能的大幅提高以及机器学习技术的应用和发展，许多研究者开始将基于神经网络的分类器应用于语音情感识别。其中深度神经网络（Deep Neural Network, DNN）、卷积神经网络、循环神经网络等网络模型在语音情感识别的特征提取、分类器构建上都有着广泛的应用，并取得了很好的效果。例如，R Burget 等使用深度神经网络对柏林库中的三类情感语音进行分类，将语句分为 20ms 的帧后，对每一帧进行上下文无关的分类，在测试集上取得了 77.51%的帧准确率和 96.97%的语句准确率[29]; Z Huang 等通过卷积神经网络对语谱图进行特征提取，最后通过支持向量机进行分类，并在 SAVEE, EMO-DB 等四个数据集上进行了模型验证，取得了高于传统语音情感识别方法的识别率[30]。

目前，语音情感识别在服务业、医疗、教育等领域已经得到广泛的应用，其应用场景也在不断扩充，具有广阔的市场前景。总而言之，在人机交互中至关重要的语音情感识别必将在即将到来的人工智能时代中发挥巨大的作用。

上述为常见的单维语音信息识别技术。对于多维语音信息识别技术，国内外研究成果很少，一种原因是每一种单维信息识别系统的性能还有提升空间，尤其是鲁棒性，人们的关切点还集中在继续完善单维系统性能上；还有一个原因可能是研究人员觉得可以将各个单维识别技术串联起来，形成多维识别结果。本团队认为，后者不符合人类识别的模式，也没有利用各单维识别任务信息之间的相关性，长远看必将构建多维信息识别系统。因此开始探索，并已进行了一段时间的研究，也产生了一些成果。例如本团队成员在文献[7]中使用支持向量机建立了一个语音多维识别基线系统，利用说话人性别相关信息提高了模型对于说话人身份和情感的分类性能，证实了不同分类任务之间信息的相关性和多维语音识别的有效性；在文

献[5][6]中分别使用多任务全连接网络和渐进式神经网络建立多维识别模型，将语音提取的i-vector特征输入到网络中，同时输出说话人性别、身份、情感三个分类任务的分类结果。上述工作也有一些局限性，例如没有很充分利用到各单维识别任务之间的相关性，特征选取和网络模型也较为简单等。

除此之外，目前国内外并没有很多关于多维语音识别的研究，但有一些概念相关的研究。例如文献[31]从多维信息识别的角度研究了人脸识别，这与多维语音识别有着相同的思想。在语音识别领域，一些研究将多任务学习和语音识别相结合，例如文献[28]构建了一个具有两层隐层的多任务全连接网络对三种情绪属性（Arousal, Valence, Dominance）进行同时分类，在一定程度上利用了各任务之间的共有信息；文献[32]基于深度置信网络（Deep Belief Network, DBN）提出了一个多任务学习框架，主任务为情感类别，辅助任务为两种情感属性activation 和 valence 的离散分类或线性预测，最后通过支持向量机进行分类，取得了高于单独情感识别的识别率。此外，文献[94]中提出的多示例多标记学习（Multi-Instance Multi-Label learning, MIML）框架考虑到输入和输出空间的多义性，即同一个对象可以有不同的特征以及属于不同的类别，同一种类别中可以包含不同的对象。此思想可以引申到多维语音识别上，利用标记之间的相关性进行多维判决。本团队成员在文献[7]中即做了相关研究，利用 MIML 算法和支持向量机建立了基于性别的身份、情感双重判决模型，取得了优于单维识别的识别率。以上研究从不同的角度证实了多个维度的语音信息同时识别是可行且有效的。

1.3 多维语音信息识别技术

目前，研究者们对各类单维语音信息识别的研究已经持续了大半个世纪，在取得了许多进展的同时也存在很多还未解决的问题，离实现真正的人机语音交互还有很长的距离。单维语音信息识别技术目前存在的问题主要表现为以下几点：

（1）实验规模较小。目前的单维语音信息识别技术的研究，往往针对所研究的单项识别任务采用对应合适的语音库，往往规模较小，针对性强，且没有考虑到实际环境中复杂的语音场景。以此训练出的模型虽然有不错的性能，但在应对大数据时代下的大型语音数据库就显得捉襟见肘，在实际应用中也有许多局限性。

（2）功能单一，模型重复。对于识别多种语音信息的任务需求，只能分别采用对应的单项语音识别技术，对于同样的语音样本需要多次可能重复的处理，存在效率低下的问题。

（3）语音信号中包含的信息没有充分有效的利用。语音信号中包含许多种类的信息，不同语音信息之间存在相关性，是互相影响的，在特征提取时，很难针对某一任务提取与其相

关且有效的特征参数。另外，对于不同任务之间的共性特征和差异性特征，也无法明确的区分和分离。

总而言之，目前的单维语音信息识别技术主要关注单一语音信息的识别任务，还无法对语音中包含的丰富信息进行全面、深度的挖掘，距离最终的拟人化语音交互还很遥远。单维识别往往只考虑任务相关的特征信息序列，对于其他丰富的语音信息例如各识别任务之间的共有信息则直接摒弃，这显然不利于理解语音的真正含义，也不是人机语音交互正确的发展方向。因此，本团队针对目前语音识别的研究现状，提出了这个多维语音识别的课题，试图充分利用语音信号中的各类信息，深入挖掘多维语音信息之间的关联性，提取各识别任务的共有特征，提升语音识别的性能，使得语音识别向拟人化更进一步。

多维语音识别指的是一个识别模型可以从输入的语音中同时识别出多种语音信息或说话人信息，多个识别任务互相利用任务间的相关信息，像人脑一样同时产生多项识别结果。相较于单维语音识别来说，多维语音识别可以更好的利用到语音中所包含的不同维度的信息，有更好的识别性能，也更加“拟人化”，更符合未来人机语音交互的场景。此外，随着人工智能和大数据技术的发展，多维语音信息识别技术也能够更好的应对庞大的数据量，符合未来的发展趋势。多维语音识别有两个主要难点，分别是多维特征参数的提取和多维识别模型的构建。经过对单维语音识别技术的学习和研究，本文在特征参数提取方面分别采用了 MFCC 和卷积神经网络提取的特征，在模型构建方面结合了多任务学习、循环神经网络、卷积神经网络来构建多维识别模型。

1.4 论文内容及论文安排

本课题基于导师承担的国家“863”高技术研究发展计划项目（“多语言语音识别关键技术研究与应用产品开发”，编号：2006AA010102）以及国家自然科学基金项目（“鲁棒性压缩感知关键技术的研究”编号：61271335），是这些语音处理国家项目后续的研究拓展。课题要求对语音进行特征提取，建立分类模型同时对语音中的多维信息进行识别；采用多任务学习理论，选取相关识别任务，提取合适的特征，建立分类模型，优化系统参数，最后输出语音的多维识别任务的分类结果。本文主要研究的是说话人性别、情感、身份三种维度信息的联合识别，利用三种识别任务之间的相关信息，取得更好的识别性能。未来其基本原理通过继续改进和完善，可以进一步推广到更多的多维信息同时识别任务中。

全文共分为五章内容，其中各个章节的主要内容如下：

第一章：绪论。本章主要介绍了多维语音识别技术的研究背景和研究意义。介绍了国内

外对于说话人身份识别、说话人性别识别、说话人情感识别三种单维语音识别的研究现状以及国内外和本团队对于多维语音识别的研究进展；接着通过分析单维语音识别存在的问题，介绍了研究多维语音识别技术的意义。最后概括性的描述了本文的章节内容安排。

第二章：多维语音识别技术概述。本章主要介绍了多维语音识别技术所涉及到的各项关键技术。首先介绍了语音的产生过程和声学模型，阐述了单维和多维语音识别的主流流程；接着介绍了语音信号常用的几种预处理方法以及常用的语音特征参数的提取；最后介绍了常用的单维语音识别模型以及多维语音识别模型的构建。

第三章：研究基于多任务学习和循环神经网络的多维语音识别。本章首先介绍了通过多任务学习实现多维语音识别的有效性以及循环神经网络在处理时序数据的优越性，接着构建了基于多任务学习和循环神经网络的多维语音识别模型，将 MFCC 特征输入到模型中，同时输出说话人身份、性别、情感三项识别任务的分类结果。算法模型通过 RNN 共享层共享网络参数以学习各任务间的共享特征，通过全连接属性依赖层学习独有特征，从而提升分类性能。还可以通过调整损失函数中各任务的权重，来针对不同数据库进行模型性能的优化。最后，进行模型仿真实验以及各项对比实验，实验结果表明，本章提出的多维识别模型在各项识别任务上相对于单维识别模型均有一定程度的性能提升，且对于语种和说话人的个性因素有较好的鲁棒性，并具有一定的抗噪性能。

第四章：研究基于卷积神经网络和特征融合的多维语音识别。本章首先介绍了卷积神经网络在语音特征提取上的应用以及优越性，构建了基于卷积神经网络的端到端多维语音识别模型，将语谱图输入 CNN 中进行特征提取，提取出的特征输入到多任务网络中进行分类，最后同时输出说话人身份、性别、情感三项识别任务的分类结果。接下来分析了单一特征参数的不足，介绍了特征融合的方法，构建了基于卷积神经网络和特征融合的多维语音识别模型，将 CNN 提取的特征与 MFCC 进行特征融合，随后输入到 RNN 共享层和全连接属性依赖层中进行多任务的同时分类。最后进行模型仿真实验和各项对比实验，实验结果表明基于融合特征的多维语音识别模型有效地提升了多维语音识别的性能，且具有更好的抗噪性能。

第五章：总结与展望。对全文的研究工作成果和不足进行总结，对多维语音识别技术的未来发展前景和研究方向进行展望。

第二章 多维语音识别技术概述

本课题的研究内容是多维语音识别技术，想要搭建科学有效的多维语音识别模型，首先要对现有单维语音识别技术框架的流程、模型、技术细节等有深入的了解。通过对现有单维语音识别方法的研究和分析，找出各识别任务之间的相关性和差异性，为多维语音识别研究提供经验、寻找可借鉴之处，在此基础上进行多维语音识别模型的设计。本章主要介绍多维语音识别技术所涉及到的一些基础知识和关键技术。

2.1 语音声学模型

要研究语音识别，首先需要弄清楚语音的声学模型，即语音是如何从人类的发声器官中产生，以及如何对此进行数学建模，以方便对语音信号进行科学的处理。

在发音语音学中定义的人体发音器官包括肺和声带、气管、咽喉、口腔鼻腔、嘴唇等[1]。语音的产生是由大脑通过向神经中枢发送包含语音信息的信号，控制上述发声器官发出声音，其中，从鼻腔中发出的声音叫鼻音，从口腔中发出的声音叫口音。发声器官的发声首先由肺部产生气流，经过喉咙处的声带和声门产生气波，经过气管、口腔和鼻腔，发生共振最终发出语音。其中，声带不振动发出的声音叫做清音，声带振动发出的声音叫做浊音。

随着对语音信号研究的深入，人们开始尝试对语音信号进行建模。研究表明，人类的语音主要有两种特征，第一是频率集中于 0.3kHz 到 4kHz 之间，第二是时变但具有短时平稳特性[33]。第一点使得对语音的采样率必须大于 8Khz，第二点使得对语音信号的处理可以通过分割为持续时间较短的帧来进行。

根据语音信号的特性，研究者们抽象出了各种声学模型，较为常用的有声管模型、线性预测编码（Linear Predictive Coding, LPC）模型和共振峰模型等。

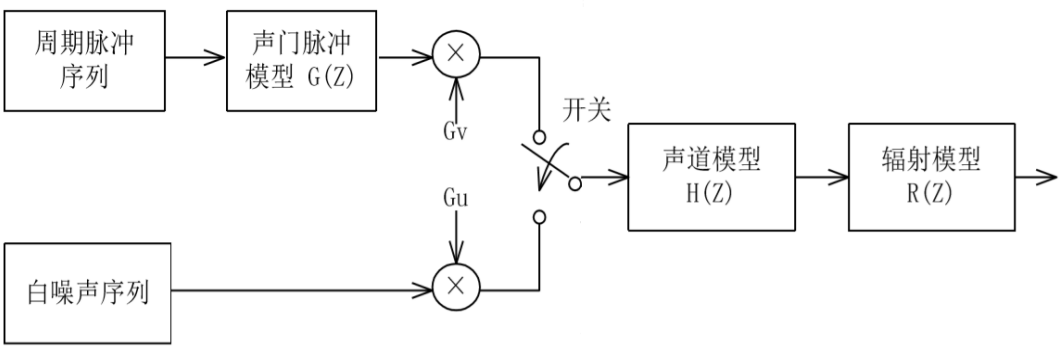

图 2.1 LPC 模型示意图

LPC 模型如图 2.1 所示，其基本原理是将语音信号 $S(z)$ 看作原始激励信号 $U(z)$ 经过声道模型 $H(z)$ 的输出，通过模型参数来对语音信号进行描述。$H(z)$ 的表达式如下所示：

$$H(z)=\frac{S(z)}{U(z)}=\frac{G}{1-\sum_{i=1}^{P}a_i z^{-i}} \tag{2.1}$$

其中 a_i 为线性预测系数，G 为系统增益，这样可以把语音信号表示为有限数量的参数。

共振峰指的是语音频谱中能量相对集中的一些区域[1]，反映了声道的物理特性。共振峰模型以共振峰频率和带宽作为参数，设置多个滤波器，模拟人体声道特性，经过辐射模型后得到语音信号。共振峰模型是一种对声道模拟较为准确的模型，因此合成语音的质量较高，在类别上可以分为级联型、并联型和混合型，图 2.2 为混合型共振峰模型的示意图。

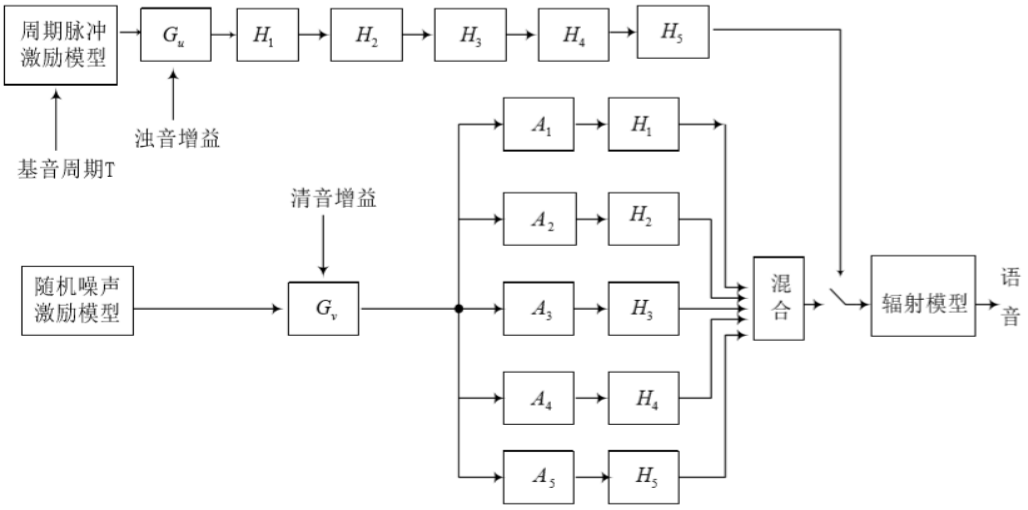

图 2.2 混合型共振峰模型[1]

2.2 语音识别主流框架

2.2.1 单维语音识别流程

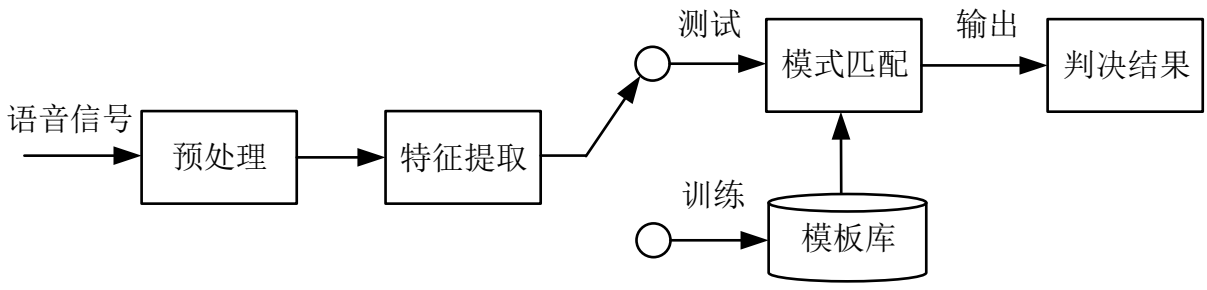

图 2.3 单维语音识别系统框图

虽然语音识别是一个很大的范畴，但对于不同类别的单维语音识别来说，语音信号从处理到识别的流程框架基本是一致的，包括语音预处理、特征提取、模型匹配、结果判决四个步骤，如图 2.3 所示。系统由训练阶段和测试阶段两部分组成[34]。在训练阶段，语音信号经过各种预处理之后，通过某种方式进行特征提取，获得含有说话人信息或语句信息且具有区分性的特征向量。随后使用该特征向量对模型进行训练，使模型参数符合训练语音中的特征。然后是测试阶段，将测试语音经过同样的处理后同训练模型进行模式匹配，从匹配结果中得到判决结果。不同语音识别任务的区别在于提取的特征和训练的模型不同，而所谓单维识别指的是模型输出的识别结果只有一个。

2.2.2 多维语音识别流程

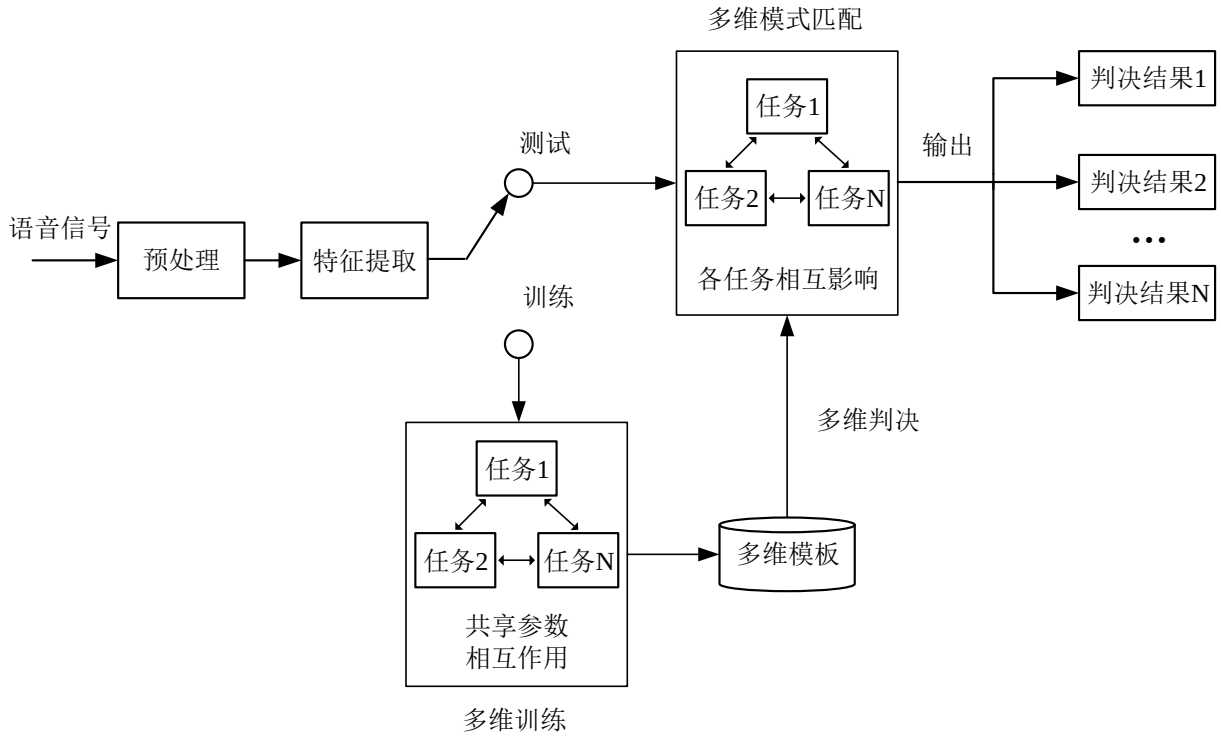

图 2.4 多维语音识别系统框图

根据单维语音识别的流程，我们可以类比设计出多维语音识别的流程框架，如图 2.4 所示，同样包括预处理、特征提取、模型匹配、结果判决四个步骤。首先多维识别有多个识别任务，系统最后同时输出多个任务的判决结果。在预处理部分，多维语音识别与单维语音识别类似，都是对语音信号进行语音活动检测、分帧、加窗、增强、降噪等，使其更加适合进行识别和分类。在特征提取部分，对于单维识别来说，不同的识别任务一般都有其所适合的

特征参数；而对于多维语音识别来说，由于有多个识别任务，各识别任务之间具有相关性，即需要利用到语音中的多个维度信息，故其与单维识别不同，需要尽可能地提取包含不同维度语音信息、适合多项分类任务的特征参数，在难度上显然要高于单维语音识别。在模式匹配和判决部分，多维语音识别与单维识别有很大不同。首先在训练阶段，多维语音识别同时针对多个识别任务训练多维模型，各识别任务之间共享模型参数、相互作用、相互影响，使多维模型能够同时符合训练语音中的多种语音信息的特征；在模式匹配和判决阶段，多维语音识别通过多维模板同时对多个识别任务进行判决，判决过程中不同参数和结果相互作用，各任务相互影响，最终同时输出多个识别任务的识别结果。多维模型对模型分类器的选择有一定的要求，可以是多个分类器的组合，也可以是同时输出多个判决结果的分类器。总体来看，多维语音识别模型的设计和单维语音识别模型有很多相似之处，但在特征选择和分类模型构建方面相对于单维语音识别来说有着更高的难度。

2.3 语音信号预处理

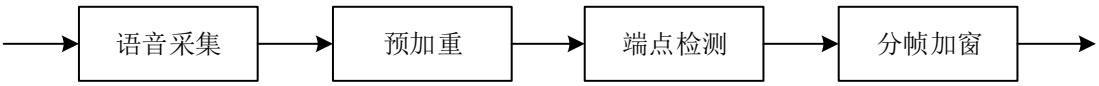

图 2.5 语音信号预处理流程

语音识别研究的第一步是对语音信号进行预处理。预处理的目的是消除语音信号从发声到采集的过程中发声器官本身以及环境噪声等对信号质量产生的影响，使其更加适合提取特征参数、进行语音识别。如图 2.5 所示，语音信号预处理的流程一般包括语音采集、预加重、端点检测、分帧加窗等，在实际中往往根据特征提取时的语音数据要求采取针对性的预处理方式，例如在语音信号背景噪声较大、干扰较多时，往往还需要在预加重之前进行语音增强和降噪等。

2.3.1 语音采集

人类发声器官发出的原始声音信号是连续的模拟信号，在对语音进行采集时，麦克风等采集设备通过采样将模拟的语音信号转化为数字信号以便存储。由上文可知，语音的采样频率需在 8kHz 以上。采样后的语音信号经过量化和编码后变为二进制数据存储在电子设备中，准备进行后续的处理。

2.3.2 语音增强

显然，在实际场景中并没有消声室中完美的语音采集条件，采集到的语音信号必然存在大量的环境背景噪声，这将在一定程度上影响语音信号的质量。语音识别的特征参数大多数是对噪声和环境相当敏感的，而由于噪声混入等原因，训练时的场景与实际应用场景往往有所不同，造成特征参数变化，导致识别失败。因此，背景噪声较大会对信号处理分析和研究产生较大影响，往往需要进行语音增强和消噪等措施。

语音增强的原理是根据语音信号和噪声信号本身存在的差异，直接消除语音信号中的噪声，或降低噪声信号对声音信号的影响，提取纯净的原始语音信号[35]。语音降噪的方法包括谱减法、参数谱估计法、维纳滤波、子空间法、盲信号分离法、小波变换等[36]。本团队也曾研究过基于压缩感知的语音增强技术，并取得了不错的成果[37]。

语音识别技术的最大难点在于语音本身的非平稳性以及实际场景与训练场景不同的变化影响，这也是影响语音识别产品全面应用的主要因素，所以提高系统鲁棒性，始终是语音识别领域的任务。因此在很多场景下，语音消噪是非常重要的预处理操作。

2.3.3 预加重

由于人类发声器官的特性，语音信号在高频率的分量较弱，容易被噪声所掩盖。为了对高频部分进行补偿和增强，往往需要对语音信号进行预加重，使语音信号变得更加平滑，方便进行后续的处理。预加重的方式是使经过采样量化后的语音信号通过一个高通滤波器，滤波器的阶数为 1，如下式所示。

$$H(z) = 1\text{-}\alpha z^{-1} \tag{2.2}$$

其中 α 为预加重系数，一般设置为 0.9 到 1 之间。语音信号通过滤波器的表达式如下。

$$y(n) = x(n) - \alpha x(n-1) \tag{2.3}$$

$y(n)$ 为经过预加重后的信号。

2.3.4 端点检测

端点检测的全称为语音活动性检测（Voice Activity Detection, VAD）[38]。由于人类发声器官是通过呼吸来发声的，导致语音信号在时域存在断续，而端点检测的作用就是找到语音信号部分的起始点和终止点，从而去掉语音信号不存在的静音部分，获取语音信号中的有效语

音。

对信号进行端点检测的方法有很多，在语音信号处理领域常用的方法叫做“双门限法”[39]，双门限的意思是基于短时能量门限和短时过零率门限来判决有用语音信号。首先要计算语音信号的短时能量，计算方法如下式：

$$E_i = \sum_{m=0}^{N-1} [x_i(m)]^2 \tag{2.4}$$

其中 $x_i(m)$ 为语音信号的第 i 帧。由于语音帧和噪声帧的能量是不同的，所以可以通过设定高低两个短时能量门限来判决出语音信号的起始点和终止点。但是，由于清辅音的能量也很低，且当语音信号信噪比很低时，通过短时能量来判决的误判率很高，仅仅通过短时能量来进行端点检测是不够严谨的。所以需要在短时能量判决之后再通过短时过零率进行二次判决。语音信号平均短时过零率的计算方法如下式：

$$Z_i = \frac{1}{2} \sum_{m=0}^{N-1} |\operatorname{sgn}[x_i(m)] - \operatorname{sgn}[x_i(m-1)]| \tag{2.5}$$

其中 N 为帧数。通过设置高低两个过零率门限来判断出语音信号的起始点和终止点，得到有效的语音信号。

2.3.5 分帧加窗

语音信号是一种具有时变特性的非平稳随机过程[1]，因此，对其进行建模是非常困难的，但语音信号存在“短时平稳”的特性。由于人体发声器官是通过神经中枢控制肌肉组织发出声音，而肌肉组织的运动是相对缓慢的，故语音信号在非常短的时间内没有剧烈的变化，可以看作近似平稳过程。因此，在语音信号处理时常常将其切割为许多短时帧，把每一帧看作平稳过程进行分析处理。一帧的长度通常取 10-30ms。在对语音信号进行分帧时，往往相邻帧之间取部分重叠，如图 2.6 所示，其目的是避免相邻帧之间出现信息断裂、不连贯的情况[40]。

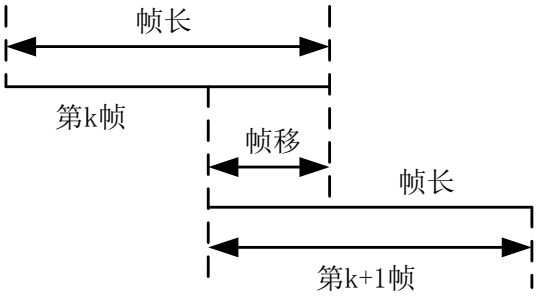

图 2.6 语音信号的交叠分帧法

在分帧之后，还需要对信号进行加窗处理。加窗的目的是使得分帧后的语音信号保证连续性，避免出现吉布斯效应[1]。语音信号经过加窗处理之后的表达式为：

$$s_{\omega}(n) = s(n) * \omega(n) \tag{2.6}$$

其中 $\omega(n)$ 为窗函数，常用的窗函数有如下几种：

（1） 汉明窗

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), & 0 \le n \le (N-1) \\ 0, & \text{其他} \end{cases} \tag{2.7}$$

（2） 矩形窗

$$\omega(n) = \begin{cases} 1, & 0 \le n \le (N-1) \\ 0, & \text{其他} \end{cases} \tag{2.8}$$

（3） 汉宁窗

$$\omega(n) = 0.5[1 - \cos(\frac{2\pi n}{N-1})], \quad 0 \le n \le (N-1) \tag{2.9}$$

N 表示窗函数的长度。其中汉明窗的旁瓣峰值衰减较小，是最常用的窗函数。

2.4 语音特征参数提取

语音信号经过预处理之后进行特征参数提取，这是语音识别流程中非常关键的一个步骤。语音信号特征参数提取指的是从语音信号中提取出能够表示该识别对象某些信息、区别于其他对象的独特特征。如上文所述，语音信号实际上是混合信号，包含许多种类的信息，特征提取即是去除语音信号中的冗余信息，提取出目标语音识别任务所需的语音特征信息，使得分类模型可以更好的进行分类任务。在单维语音识别系统中，往往根据所识别的单维语音信息设计合适的特征提取方法，然后在分类模型中使用所提取的特征参数进行判决和识别。

为了更好的进行多维语音识别的语音特征参数提取，需要先对单维语音识别中常用的语音特征参数进行总结分析。对于不同的语音识别分类任务，往往选择的特征参数有所不同。目前，已经有非常多的语音特征参数提取方法，大体上可以分为三个类别：韵律特征、谱特征、其他特征。接下来分别对三种特征中的典型特征进行介绍。

2.4.1 韵律特征

韵律特征指的是语音信号在文本无关的音高、音长、音强、快慢等方面的结构性信息，

又被称为“超音段特征”或“超语言学特征”，可以分为三个方面：语调、时域分布和重音，是语言和情绪表达的重要形式之一[24]，经常被用于语音情感识别的研究中。常用的韵律学特征包括但不限于以下几种：

（1） 能量

语音信号的能量反映了信号的强度，在实际中体现为声音的强弱和轻重音，由信号的幅度决定。一帧语音信号的能量计算如下式所示：

$$E_n = \sum_{m=-\infty}^{\infty} \left[x(m)\omega(n-m)\right]^2 = \sum_{m=n-N+1}^{n} \left[x(m)\omega(n-m)\right]^2 \tag{2.10}$$

由能量可以衍生出能量轮廓，并计算方差、均值、范围等统计学特征参数[41]。

（2） 时长

语音信号的时长表示语音的长度，是语音信号的时间特性，包括说话时间、静音段比例等，反映了语音信号的基础结构信息[41]，可以构成有区别作用的“长短音”。

（3） 基音频率

语音信号的基音频率反映了语音的音高，表示的是人类发声器官中的声带振动的频率，构成了语音的音调。由基音频率可以衍生出基频轮廓，计算其差分、方差、均值、范围等统计学参数作为特征参数[41]。

对基音频率进行检测的方法有很多，较为常用的是自相关基音周期检测法[1]。由于语音信号是周期信号，由信号处理知识可知其短时自相关函数具有相同的周期，通过计算自相关函数的峰值点就可以得到基音周期。具体步骤如下：首先通过“中心削波”等方式消除信号低频部分的干扰，减少无关信号信息，削波函数如下式：

$$y(n) = C\left[x(n)\right] = \begin{cases} x(n) - C_L, & x(n) > C_L \\ 0, & \left|x(n)\right| \le C_L \\ x(n) + C_L, & x(n) < -C_L \end{cases} \tag{2.11}$$

其中 C_L 为削波电平，一般取最大幅度的百分之六十到七十。削波函数波形如图 2.7 所示

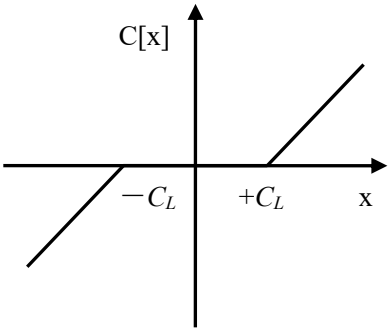

图 2.7 中心削波函数

信号经过削波后通过其短时自相关函数计算得相应的基音周期，继而得到语音信号的基音频率。信号短时自相关函数的表达式如下：

$$R_n(k)=\sum_{m=n}^{n+N-k-1}x_\omega(m)x_\omega(m+k) \tag{2.12}$$

在实际研究中，往往以语音信号的能量、时长、基音频率等为基础进行数学计算、统计分析和特征选择，得到所需的语音特征参数[24]。

2.4.2 谱特征

与韵律特征的连续性不同，谱特征通常是语音信号的短时表示[42]。目前对于谱特征的应用非常广泛，诸如 LPCC、MFCC 等在各类语音识别任务中都有着非常好的性能表现。

（1） 线性预测倒谱系数（LPCC）

由于语音信号存在相关性，故某时刻的语音信号采样值可以近似表示为已知的、过去的语音信号采样值的线性表示，该方法叫做线性预测（Linear Prediction, LP）[13]。当预测值与实际值误差最小时的预测系数叫做线性预测系数（Linear Prediction Coefficient, LPC）[44]，该系数可以作为语音信号的一种特征参数，可以反映出声道的变化。预测值的计算如下式所示：

$$\hat{s}(n)=\sum_{i-1}^{p}\lambda_i s(n-i) \tag{2.13}$$

其中，λ_i 表示线性预测系数。因此我们可以计算出预测误差：

$$e(n)=s(n)-\hat{s}(n)=s(n)-\sum_{i-1}^{p}\lambda_i s(n-i) \tag{2.14}$$

当误差最小时，要求满足式 $\frac{\partial E[e^2(n)]}{\partial\lambda_i}=0$，可以得到下式：

$$R_x(j)=-\sum_{k=1}^{p}\lambda_k R_x(i-j),\qquad i=1,2,...,p \tag{2.15}$$

将方程展开为矩阵形式，可得：

$$\begin{bmatrix}R_x(0) & R_x(1) & \cdots & R_x(p-1)\\ R_x(1) & R_x(0) & \cdots & R_x(p-2)\\ \vdots & \vdots & \vdots & \vdots\\ R_x(p-1) & R_x(p-2) & \cdots & R_x(0)\end{bmatrix}\begin{bmatrix}\lambda_1\\ \lambda_2\\ \vdots\\ \lambda_p\end{bmatrix}=\begin{bmatrix}R(1)\\ R(2)\\ \vdots\\ R(p)\end{bmatrix} \tag{2.16}$$

上述方程的解 $\{a_i\}$ 即为所求的 LPC 系数。

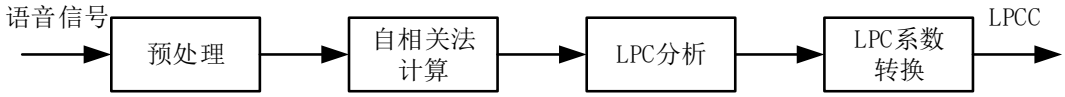

图 2.8 LPCC 提取过程

在线性预测系数的基础上进行转换，可得到线性预测倒谱系数（LPCC）。LPCC 是一种在语音识别中很常用的谱特征参数，它能够在 LPC 的基础上效去掉激励源信息，加强语音的声道特性[13]。LPCC 的流程图如图 2.8 所示，由 LPC 系数计算得到的表达式 $c(n)$ 如下所示

$$\begin{cases} c(1)=\lambda_1 \\ c(n)=\lambda_n+\sum_{i=1}^{n-1}(1-i/n)\lambda_i c(n-i), 1\le n\le p \end{cases} \tag{2.17}$$

（2） Mel 频率倒谱系数（MFCC）

Mel 频率倒谱系数是目前语音识别领域使用最广的一种语音特征参数，其描述的是语音短时功率谱的包络，在 1980 年由 Davis 和 Mermelstein 提出[45]。首先，将语音信号通过傅里叶变换后转换为语谱图（Spectrogram）。语谱图是语音信号进行傅里叶变换之后的三维图像，综合了信号时域波形和频谱图，表示语音信号的频谱随时间的变化，其中 x 轴为时间，y 轴为频率，z 轴为幅度，通过颜色灰度表示语音信号幅度的高低。为了提取出语音频谱的包络，对信号的频谱幅度取对数后取逆傅里叶变换，得到信号的倒谱（cepstrum）[1]。另一方面，由于人耳的听觉特性，即人耳对听到的声音信号的频率从低到高呈非线性变化，研究者们为了模拟人耳引入了以 Mel 为单位的音调的概念[46]。由频率转化为 Mel 频率的方法如下：

$$Mel(f)=2595*\log_{10}(1+f/700) \tag{2.18}$$

人耳能够感知到的频率转化为 Mel 频率之后基本上是均匀分布的，因此 Mel 频率非常适合用作模拟人耳的语音信号处理。MFCC 即是先将语音信号频谱转化为 Mel 频谱，再通过计算转化为倒谱，得到的倒谱系数即为 Mel 频率倒谱系数。MFCC 特征提取的流程如图 2.9 所示。

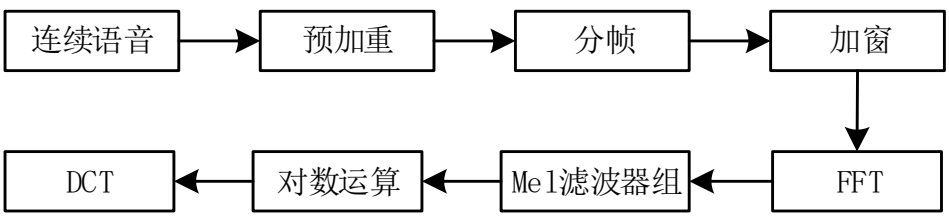

图 2.9 MFCC 特征提取流程图

信号经过 FFT 之后，取三角滤波器组作为 Mel 滤波器组，如图 2.10 所示

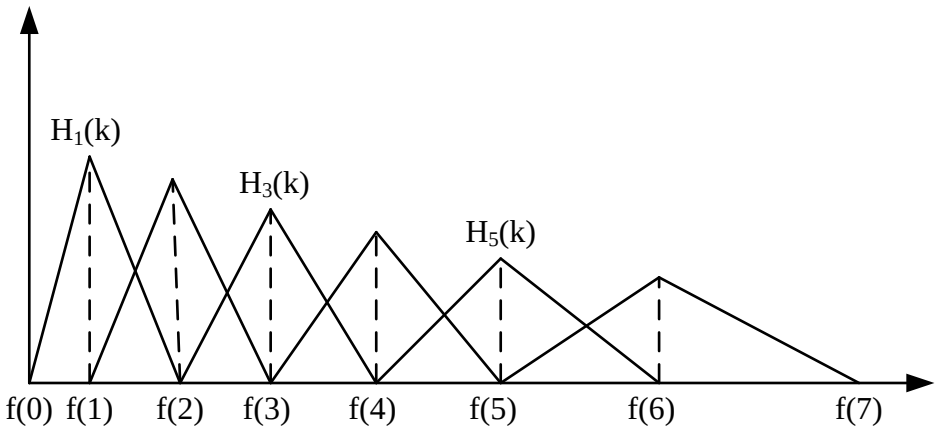

图 2.10 Mel 滤波器组

Mel 滤波器组满足每个滤波器的起点频率和相邻滤波器的中心频率相同，各滤波器的表达式如下式：

$$H_m(k)=\begin{cases}0 & ,k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m-1)\le k\le f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m)\le k\le f(m+1)\\ 0 & ,k>f(m+1)\end{cases} \tag{2.19}$$

其中，$f(\mathrm{m})$ 为滤波器的中心频率，$\sum_{m=0}^{M-1}H_m(k)=1$。

接着对滤波器的输出取对数，求得对数能量，如下式：

$$e(m)=\ln\left(\sum_{m=0}^{N-1}|X_a(k)|^2\,H_m(k)\right),0\le m\le M \tag{2.20}$$

最后进行 DCT 变换，得到 MFCC 特征，如下式：

$$C_n=\sum_{k=1}^{M}\cos\left(\frac{\pi n(k-0.5)}{M}\right)\ln e(m),n=1,2,\cdots,\ \mathrm{L} \tag{2.21}$$

其中，L 表示 MFCC 特征的维度，一般取值为 12 到 16 之间。

在实际应用中，为了达到更好的效果，往往还需计算 MFCC 特征的差分，一般把 MFCC 系数及其一阶差分和二阶差分共同作为 MFCC 特征，供后续的语音识别使用。

2.4.3 其他特征

除了韵律特征和谱特征之外，还有其他种类的特征，例如近年来在说话人身份识别领域表现优异的 i-vector 特征[48]，以及伴随着人工智能的发展出现的各类基于深度学习和神经网

络的语音特征[49] [50]等等。

（1） i-vector 特征

身份矢量（Identity-Vector）特征是基于 MFCC 特征和 GMM-UBM（Gaussian Mixture Model-Universal Background Model，高斯混合模型-通用背景模型）框架的语句特征[47]。如今的说话人身份识别领域中，i-vector 特征及其衍生特征被越来越多的研究者使用，并取得了出众的识别效果。i-vector 特征的提取流程如图 2.11 所示，首先对语音信号进行 MFCC 特征参数提取，随后在 GMM-UBM 框架下通过 UBM 模型的自适应得到说话人模型，得到蕴含说话人信息的 GMM 均值矢量，堆叠之后形成高维的均值超矢量。接着在超矢量空间中，通过训练将超矢量 M 进行分解，如下式所示：

$$M=m+T\omega \tag{2.22}$$

其中 m 表示说话人与信道独立的 UBM 均值超矢量，T 为全局差异空间，ω 为全局差异因子，先验服从标准正态分布，其后验均值即为 I-vector 矢量[48]。式中的 M 和 m 是可知的，故需要对矩阵 T 和 ω 进行估计。通过计算对应的 Baum-Welch 统计量，采用 EM 迭代算法对 T 矩阵进行迭代估计，然后将得到的 T 矩阵带入式中计算出 ω 的后验均值，即为所求的 i-vector 特征。

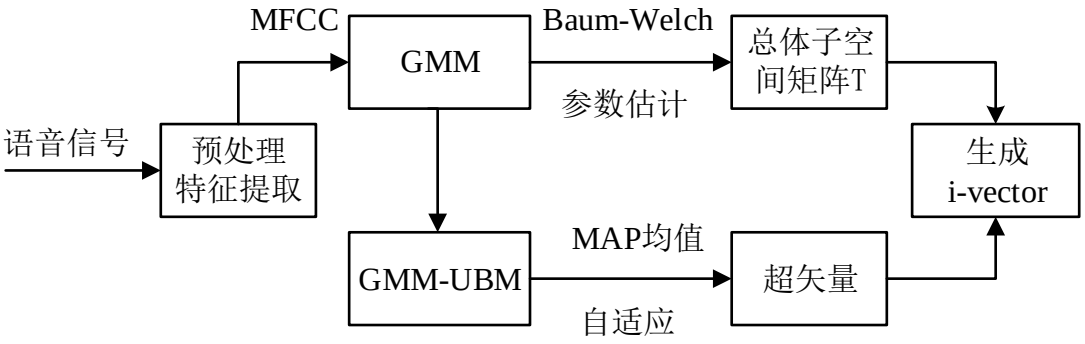

图 2.11 i-vector 提取流程图

（2） 基于神经网络的特征

近年来，随着人工智能和深度学习的不断发展，各类神经网络在语音识别领域有着出色的表现，在特征提取方面也不例外，不同种类的神经网络有着不同的优点和应用。DNN 能够将语音特征映射到独立空间，例如文献[49]将语谱图输入 DNN 中提取瓶颈特征（Bottleneck Features）用作语音情感识别，如图 2.12 所示。

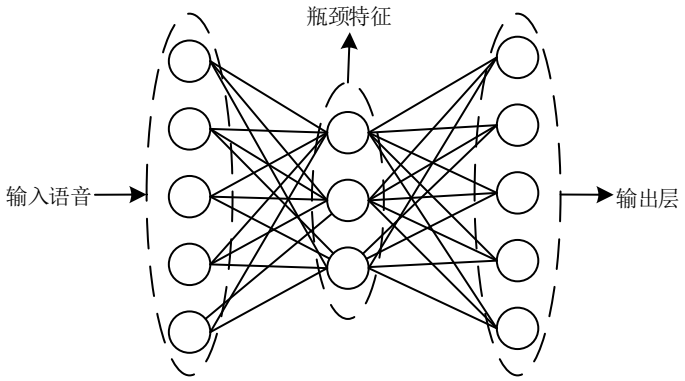

图 2.12 DNN 提取瓶颈特征[49]

RNN 中的 LSTM（long short-term memory）和 GRU（Gate Recurrent Unit）等对序列数据上下文的依赖关系具有天然的描述能力[50]，能够很好地对时序数据进行表达，适合处理语音信号，例如文献[51]的研究中使用了深度 LSTM 对语音进行处理。本文在第三章中会详细介绍 RNN 的相关内容。

CNN 可以用来进行特征降维和特征提取。CNN 在图像处理上的强大性能在语音信号处理中也有很好的应用，可以通过将语音信号转化为语谱图，以图像形式输入到 CNN 中进行特征提取[52]。也有研究使用 CNN 进行端到端的语音识别[53]。本文在第四章会详细介绍 CNN 的相关内容。

更多的语音识别研究往往将各类神经网络融合起来使用，例如知名的 CLDNN[54]网络结构，将 CNN、LSTM、DNN 融合在一起同时训练，使得到的语音特征参数在语音识别中能够取得更好的分类能力。

2.4.4 多维语音识别特征选取

上文介绍了单维语音识别中常用的特征参数。对于不同的单维识别任务，一般有其适合的特征参数。例如，在说话人性别识别中，由于其难度较低，可选的特征较多，实际的研究中多数文献采用基音频率等韵律特征结合 MFCC 进行识别；在说话人情感识别的研究中往往使用短时能量、基音频率、MFCC 等特征参数；在说话人身份识别上，i-vector 特征有着远超其他特征的性能表现。整体来看，在目前的单维语音识别研究中，往往针对所进行的单一识别任务提取单维语音信息，仅仅关注识别任务本身，这样使得特征提取过程中语音信号中存在的其他丰富的多维信息被摒弃掉。

多维语音识别要求充分利用语音信号中的各类信息，深入挖掘多维语音信息之间的关联性，因此在特征参数提取方面，多维语音识别对语音特征参数的要求比单维语音识别更加严格，显然不能直接采用经验性的单维语音识别特征来进行多维语音识别。多维语音识别所用

的特征参数需要尽可能的表达多维度的语音信息，既能同时满足多个识别任务，又能充分包含不同语音识别分类任务之间的共有信息，使识别模型能够充分学习到各识别任务之间的共有特征以及各自本身的私有特征。另外，还需要考虑到特征提取的算法效率和特征冗余度，语音特征参数中的信息需要足够丰富，但又要保证有效性，避免冗余。

考虑到以上要求，多维语音识别特征参数提取的方法主要有以下几种思路：

（1）将多维语音识别中各识别任务所适合的特征进行组合，得到组合特征。本团队成员先前在文献[2]中提出的多维说话人信息识别方法即采用了这种办法来对说话人身份、情感、性别进行识别，其使用的组合特征包含了基音频率、短时能量、共振峰、MFCC、响度等多种特征及其统计值等。但是，以这种方式得到的多维特征在信息丰富度和特征的维数之间较难平衡，容易得到冗余度较高的高维特征，且组合特征中的某类特征可能对某种识别任务起到积极作用，但可能对另外一种识别任务存在干扰，从而影响多维识别的效果。

（2）采用多维语音识别中各识别任务均表现良好的共有的特征参数。本团队成员先前在文献[3]中提出的多维说话人信息识别系统即采用了这种方法，其进行的是说话人身份、情感、性别的多维识别，分析得到三种识别任务均可使用 i-vector 特征进行识别，故采用了该特征作为多维识别系统的特征参数。这种多维特征选取方式受多维识别任务类别的影响较大，在大多数情况下难以找到各识别任务均适合的特征，且往往对于某些识别任务来说并非最优。

（3）通过其他方式提取同时适合多个识别任务的特征。使用某种工具或算法，例如卷积神经网络等，针对多维识别的多个识别任务，提取同时满足各识别任务要求的特征。这种方式最为合适，但也最难实现。

经过上文对多维语音特征参数的分析，并考虑到课题要求，本文的多维语音识别研究采用的是 MFCC 特征和基于卷积神经网络的特征，具体内容将在第三章和第四章进行详细的介绍。

2.5 算法模型选取

2.5.1 单维语音识别常用算法模型

在提取出合适的语音信号特征参数之后，需要建立合适的识别模型进行分类任务，这也是语音识别流程中的关键部分。对于多维语音识别来说，需要进行多维判决、同时输出多个识别任务的分类结果，属于多分类模型，因此要求分类器具有多分类判决的能力。在单维语音识别的研究中对于识别模型和分类器已经有了很多的进展，通过对单维语音识别算法模型

的学习和研究，可以启发出多维语音识别分类模型的构建方法。

早期常用的算法是基于模式匹配的算法，例如动态时间规整算法，其基于动态规划（Dynamic programming）思想，解决了语音长短不一的模板匹配问题，通过衡量语音序列之间的距离，进行语音识别分类任务。在后来的研究中，以 GMM、HMM 为代表的概率模型算法开始被广泛应用。GMM 通过对特征参数的概率分布进行建模代替特征参数本身，将多个高斯分布进行线性组合表示多维特征矢量的概率分布，通过判别似然得分代替语音模板距离，可以对待识别语音进行更好的描述[39]。HMM 对语音信号的时序结构建立统计模型，将其看作一个数学上的双重随机过程：一个是用具有限状态数的马尔可夫链来模拟语音信号统计特性变化的不可见的随机过程，另一个是与马尔可夫链的每一个状态相关联的外界可见的观测序列的随机过程（一般为语音特征参数）[34]。HMM 将上层的语言模型和底层的声学模型很好地融合统一，因而成为语音识别领域的重要模型得到广泛应用[2]。

近年来，由于机器学习技术的兴起和发展，许多基于机器学习的分类模型开始兴起。支持向量机是一种常用的基于机器学习的分类器。SVM 可以将低维空间的非线性不可分问题转化为高维空间的线性可分问题，其中心思想是在高维空间中寻找满足类别间距离最大条件的最优分离超平面，以此将不同类别的数据集分割开来[57]，因此可以很好的通过语音特征参数进行分类。在深度学习出现之前，SVM 被认为是机器学期近十几年来表现最好的算法。而随着人工智能的发展，计算机性能大幅度提升，各类神经网络在构建分类模型方面也有了广泛的应用。通过训练神经网络分类器，使神经网络自主学习语音特征，输出分类结果，可以充分利用到计算机性能以及大数据支撑的优势。其中诸如 DNN、RNN、CNN 等不同的神经网络有着不同的作用和侧重点，在这方面的研究和成果层出不穷，成为当前语音识别研究领域的主流。

2.5.2 多维语音识别算法模型选取

上述分类模型目前往往都是针对单维语音识别研究，而多维语音识别在单维语音识别的基础上有更多的要求。一方面，单维语音识别模型的输出只有一个任务的分类结果。对于多维语音识别，模型要求实现多任务分类，同时输出多个识别任务的分类结果，可以通过组合多个分类器来实现，也可以直接采用同时输出多个判决结果的分类器。另一方面，单维语音识别模型只需要从语音特征参数中学习到单一识别任务的特征信息，在判决时只需要进行单维判决，而多维语音识别模型需要同时学习多个识别任务的特征，其中包括各识别任务本身的私有特征以及任务间的共有特征，在判决时需要进行多识别任务同时判决，各任务之间相

互影响，得到最终的多维识别结果。

本团队前期在多维语音识别领域已经进行过一些研究，采用了一些方法来构建多维识别模型。本团队成员在文献[2]中采用了多示例多标记支持向量机算法，利用不同标签之间的制约关系，构建了先进行说话人性别识别、再进行基于性别的说话人情感、身份同时识别的多维语音识别模型。该模型在一定程度上利用了各识别任务之间的相关性，实现了多维语音识别，但也存在一些缺陷，第一是模型仍存在“串联”结构，是先进行性别识别再进行情感和身份识别的，不算是严格意义上的多维信息同时识别；第二是模型在结构上较为简单，采用的支持向量机在性能上不如最近兴起的深度学习等方法。本团队成员在文献[3]中先是采用了多任务学习理论结合深度置信网络构建了多维语音识别模型，通过 DBN 共享任务间学到的信息，进行说话人身份、性别、情感三项识别任务的同时识别；另外基于 ProgNets 的结构构建了的多维语音识别系统，在性别分类的基础上将说话人情感识别和说话人身份识别之间的信息相互迁移，实现多维识别。该文献中的模型也存在一些不足，基于多任务学习的模型其网络结构较为简单，没有很充分的利用到各单维识别任务之间的相关性；基于 ProgNets 的模型也是基于性别分类基础的，不是严格意义上的多维识别，只利用到情感识别和身份识别之间的关联性。

考虑到上述研究中的算法模型存在的问题，本文需要改进的方向是更加充分地利用各识别任务之间的相关信息，采用更科学的模型结构，提升各识别任务的性能。因此，本文在团队先前研究的基础上进行改进，选择性能更好的分类模型，将多种分类模型进行组合，目的是最大程度地适应语音特征中的多维信息、符合多任务同时分类的要求。综合上述对分类模型的研究，在近年来的研究中神经网络分类器的学习和分类能力是其他分类器不可比拟的，因此本文的研究中选择使用了 DNN、RNN、CNN 相结合的神经网络模型，并结合多任务学习理念，构造多任务神经网络分类模型，进行多维语音识别研究。通过构建具有属性依赖层的多任务神经网络[28]，使模型能够对多个任务同时进行训练学习和分类，输出多个分类结果；使用适合进行二维处理分析的 CNN 对语音语谱图进行特征提取、筛选以及降维操作；使用适合时序信号处理的 RNN 来构建网络共享层，各语音识别任务在共享层中共享网络参数以学习共有特征；使用 DNN 来构建属性依赖层，各任务可通过属性依赖层来学习本任务的独有特征。该模型最后能够同时输出语音信号对应的说话人性别、说话人身份、说话人情感三个识别任务的分类结果。模型相关的具体内容将在第三第四章进行详细的介绍。

2.6 本章小结

本章主要介绍了多维语音识别技术所涉及到的基础知识和关键技术。首先简要介绍了语音的产生过程和声学模型，接着介绍了单维语音识别和多维语音识别的流程框架，包括语音预处理、特征提取、模型匹配、结果判决四个部分，然后对各部分进行了详细的介绍。首先介绍了语音信号的预处理，通过语音增强、预加重、端点检测、分帧加窗等处理方式使语音信号能够更好的进行后续的特征提取和处理。接着介绍了语音信号的特征提取，介绍了韵律特征、谱特质、其他特征，例如帧能量、基音频率、MFCC、LPCC 等特征参数，并讨论了多维语音识别适合的和本文选取的特征参数。最后介绍了多维语音识别模型的选取，分析了常用的单维语音识别模型，同时引出本文多维语音识别模型的构建方法。本章所介绍的内容主要是从单维语音识别入手，将用到的技术引申到多维语音识别研究中，为后续的研究打下基础，提供理论支撑。

第三章 基于多任务学习和循环神经网络的多维语音识别

本章主要进行多维语音识别模型的研究，将多任务学习机制与循环神经网络结构相结合，构建多维语音识别模型，充分利用语音信号中的多维度信息，实现对说话人性别、身份、情感三种语音信息的同时识别。

3.1 引言

如前文所述，人类的发声系统发出的含有说话人各种信息的语音信号，经过环境传播后到达人耳时混入了各类的环境背景声音信号，因此人类听觉系统接收到的语音信号实际上是一种混合信号，其中包含了三大类别的信息，其一是语义内容信息，即为说话人想要表达的语言文本、语种、方言等；其二是说话人相关的特征信息，包括说话人身份、性别、年龄、情感状态等；其三是环境背景声音信息，包括各种背景噪声、环境声音、其他说话人的语音等。因此，采集到的含一句话的声音中可以分析出的信息是非常丰富的。

目前，语音识别领域的研究大多针对某一特定种类语音信息进行研究，针对语音信号这样一个复杂的混合信号进行单一信息的单维语音识别、显然不利于全面理解语音信号丰富的真实含义，也不符合未来信息时代中物联网和人工智能对于人机语音交互的场景要求。并且，经过了长时间的研究，语音识别领域每个子领域例如说话人身份识别、说话人情感识别等均已经取得了不错的效果，但相关研究已经有些瓶颈，始终难以产生突破性的进展，距离拟人化和普及应用还有很大的距离。因此，本团队从拟人化语音识别的角度出发提出了本课题，针对多项语音信息识别任务进行同时识别，挖掘不同识别任务之间的共有信息，充分利用语音信号中包含的多维度信息，试图从多维语音识别这一全新的角度寻找语音识别研究目前存在问题的解决方法。

多维语音识别作为一个新的研究课题，目前还缺乏成体系、可参考的研究成果和文献，因此本团队作为该课题的先行者，先从说话人性别、说话人情感、说话人身份三维信息的语音识别研究着手。本团队在多维语音识别课题上已经有了一些成果，最初利用 SVM 分类器设计了一个性别相关的多维识别基线系统，证明了多维信息同时识别的可行性和有效性[7]；接着使用了多示例多标记学习和 SVM 实现了多维说话人信息识别，得到了比单维识别模型和基线系统更好的识别效果[2]；后来的研究在前者的基础上进行了改进，采用了深度置信网络和渐进式神经网络（Progressive neural networks, ProgNets）结合多任务学习理念来构建多维

识别模型，通过 i-vector 语音特征进行多维语音识别，也取得了不错的效果，并减低了模型的复杂度[3]。本文在团队已有研究的基础上继续进行研究，采用了更为鲁棒的 DNN、CNN、RNN 等神经网络模型和更为科学的多任务学习结构，在特征提取和分类模型构建两方面均有所改进，取得了更好的识别效果。

本章研究的工作主要在多维语音分类模型的构建上，基于多任务学习和循环神经网络进行多维语音识别的研究。在语音特征参数的选取上，本章采用多维语音识别中的各识别任务均表现良好的共有的特征参数。根据文献分析和实验验证，目前被广泛使用的 MFCC 特征在说话人性别识别、身份识别、情感识别中均有良好的表现，因此本章采用组合 MFCC 特征作为特征参数，并且在第四章的工作中侧重研究特征参数的改进。在识别模型的设计上，采用了带有属性依赖层的多任务神经网络结构，选择适合处理时序数据、表征语音信号特征的 RNN 作为共享层，选择全连接网络作为属性依赖层，组合构建了多任务神经网络分类模型，输入 MFCC 组合特征，最后同时输出说话人性别、情感、身份三项识别任务的结果。随后对本章提出的模型进行仿真实验，通过设置多个对比实验，验证本章模型的有效性和先进性。实验表明，本章提出的基于多任务学习和循环神经网络的多维语音识别模型对于三项识别任务均有一定程度的性能改善，其识别准确率相对于本团队先前的研究有较大的提升，对于语音库、语种和说话人的个性因素有较好的鲁棒性，且具备一定的抗噪性能。

3.2 循环神经网络

3.2.1 循环神经网络概念

循环神经网络是一种隐层具有自连接关系的神经网络，网络的自连接特性使其对序列数据上下文的依赖关系具有天然的描述能力[50]。在全连接神经网络和卷积神经网络中，层与层之间是全连接或部分连接的，但每一层内的节点间是无连接的，造成对连续序列数据的处理能力有限。而循环神经网络通过自连接来记忆之前的信息，并对后面节点的输出产生影响，因此在处理时序数据时比其他类型的神经网络例如全连接网络和卷积神经网络，有着更好的表达能力。如今，循环神经网络已经被广泛应用于各类语音识别、图像识别、自然语言处理等多个领域。因为语音最成功的模型就是 LPC 模型，本质上是一个时变的 AR 模型，带循环反馈成分，故 RNN 天然比较适合语音处理。

典型的 RNN 结构如图 3.1 所示，图中的左边部分可以看作一个 RNN 节点，将其按时间进行展开后得到右边部分，其中 x_t 为 t 时刻的输入。节点 A 根据当前时刻的状态 A_t 计算出当

前时刻的输出值 h_t，而 A_t 由节点上一个时刻的状态 A_{t-1} 和当前时刻的输入 x_t 所共同决定，因此可以达到记忆之前信息的效果。在模型的训练和参数优化方面，与其他神经网络一样，RNN 同样定义损失函数、使用反向传播算法和梯度下降算法训练模型，其损失函数定义为所有时刻上损失函数的总和。

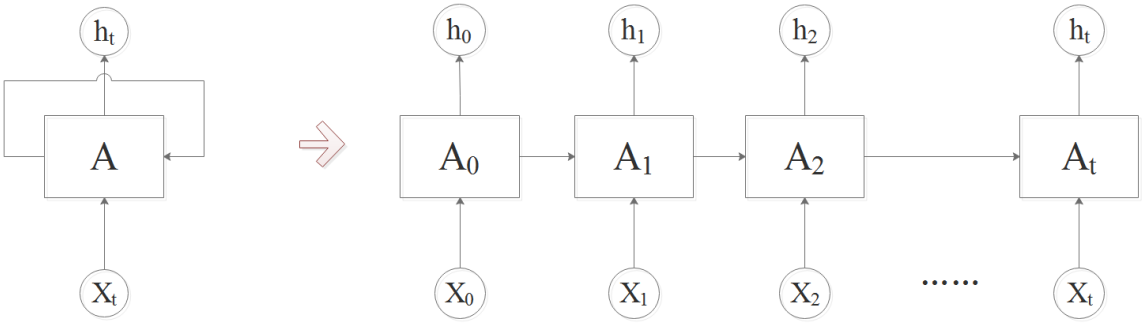

图 3.1 典型的 RNN 结构

3.2.2 LSTM 和 GRU

在普通的 RNN 模型中，由于其结构较为简单，当时域展开后步数较长时会出现“长期依赖”问题，即随着时间的延长，经过多阶段传播后的梯度将趋向于消失，造成某一时刻的输入对后续时刻的影响会越来越小，甚至无法产生影响。因此普通的 RNN 模型无法很好的表示时序数据中长距离的依赖关系[59]，这在很大程度上使得 RNN 的应用产生了局限性。而语音这样的信号，除存在短时相关性外，还存在长时相关性，即基音相关性。为了解决此类问题，除了采用一些正则化操作以及多时间尺度的策略外，往往采用 LSTM 或 GRU 网络。

LSTM 模型如图 3.2 所示。LSTM 使用三个“门”计算即遗忘门、输入门和输出门来控制不同时刻的状态和输出。每个 LSTM 节点包含三个输入，即上一时刻的节点状态 c_{t-1}、上一时刻的节点输出 h_{t-1} 和当前时刻的输入 x_t。其输入输出间的关系可由式（3.1）~（3.6）表示。

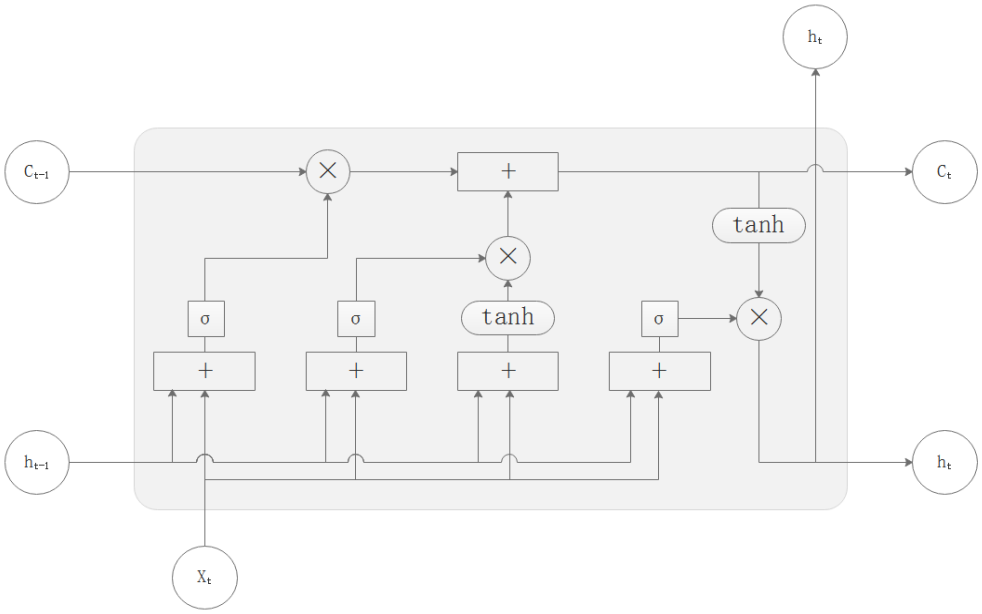

图 3.2 LSTM 节点模型

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{3.1}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{3.2}$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{3.3}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{3.4}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{3.5}$$

$$h_t = o_t * \tanh(c_t) \tag{3.6}$$

其中，σ 为激活函数，W 和 U 为权重矩阵，b 为偏置项；f_t 为忘记门，用来计算哪些信息需要舍弃，通过激活函数处理后，以一个 0 到 1 之间的值表示上一时刻的节点状态 c_{t-1} 中有多少会保留到当前时刻的节点状态 c_t；o_t 为输出门，用来计算当前节点状态 c_t 要输出多少作为节点的输出 h_t；输入门分为两部分 i_t 和 $\tilde{c}_t$，用来计算保留多少当前时刻的输入 x_t 到当前时刻的节点状态 c_t。由于三个“门”结构的存在，LSTM 在反向传播的过程中可以很好的解决梯度消失问题。

GRU 是 LSTM 的一种变体，其结构和输入输出关系式如图 3.3 和式（3.7）~（3.10）所示。GRU 和 LSTM 有类似原理，其将 LSTM 的三个门简化为两个门，分别为更新门 z_t 和重置门 r_t，x_t 和 h_t 分别表示当前时刻节点输入和当前时刻节点状态，W 和 U 为权重矩阵，b 为

偏置项，$\tilde{h}_t$ 为节点候选状态。节点状态 h_t 中使用更新门 z_t 来控制包含过去时刻信息的前一时刻状态信息 h_{t-1} 流入到当前状态 h_t 中的程度，经过激活函数 σ 处理后得到一个 0 到 1 之间的值，其值越大说明前一时刻的状态信息被带入越多；候选状态中使用重置门 r_t 来控制节点忽略前一时刻的状态信息 h_{t-1} 的程度，同样经过激活函数 σ 得到 0 到 1 之间的值，重置门的值越小说明忽略得越多，重置门近似 0 则 h_{t-1} 将被丢弃。

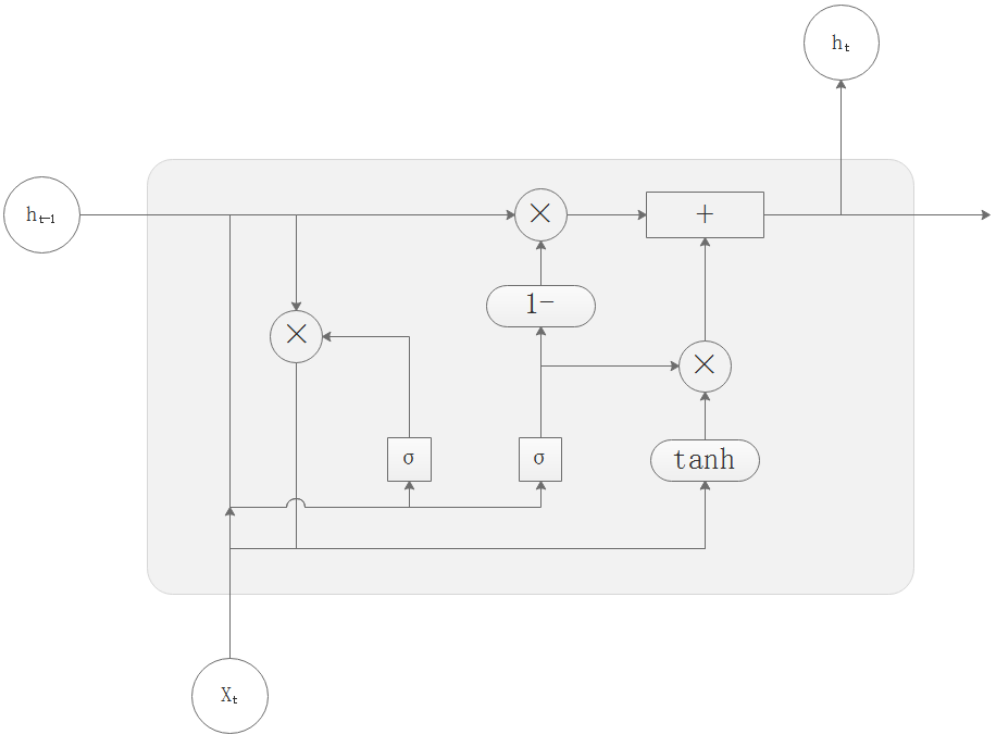

图 3.3 GRU 节点模型

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3.7}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{3.8}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h r_t * h_{t-1} + b_h) \tag{3.9}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{3.10}$$

总的来说，重置门有利于捕捉时序数据中短期的依赖关系，更新门有利于捕捉时序数据中长期的依赖关系，因此可以应对循环神经网络中的“长期依赖”问题。GRU 相对于 LSTM 来说结构更加简单，训练速度更快，参数更少因此更容易收敛，同时很大程度上保持了 LSTM 的效果。在很多训练任务上尤其是数据集不算特别大的情况下，GRU 和 LSTM 的性能几乎没有差别。

对于本文的多维语音识别研究来说，由于语音信号数据在时间上拥有很强的短时和长时等多种相关性，因此首先考虑采用 RNN 来构建识别模型可以很好地使模型学习到语音信号特征，且与多任务学习结构可以完美融合，进行多任务的学习和分类，可以有效提高模型性能。同时，考虑到语音中存在的基音长时相关，在经过分析和实验测试后，综合考虑模型性能和训练速度，本文选择 GRU 作为 RNN 的基本单元。

3.3 多任务学习

3.3.1 单任务学习

单任务学习指的是通过训练数据针对某个特定的分类任务进行独立的优化，直到使模型的结果不能继续优化即达到最佳。目前，绝大多数的机器学习和语音识别研究都是单任务学习，一次针对一个目标任务进行训练，优化模型参数。单任务学习模式下，如果要同时得到说话人性别、情感、身份三个任务的识别结果，需要为每个任务分别单独建立识别模型，分别使用数据集进行模型参数的训练和测试，如图 3.4 所示。

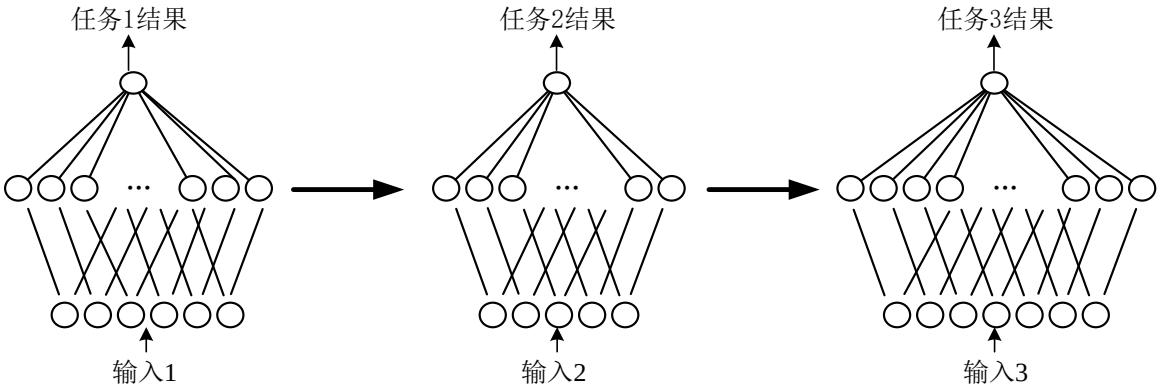

图 3.4 单任务学习模型

在传统的机器学习研究中，多任务识别往往都是采用这种分割成多个单任务模型，再进行整合的方式，各识别任务以串联的方式依次识别，最终得到多个任务的分类结果。这样的单任务学习模型，对于每个识别任务来说，虽然采集的数据相同，但是从识别特征的提取到识别模型的结构均不尽相同，每个识别任务之间相互完全独立，各模型的关注点仅仅局限于所训练的任务本身，忽略了可能帮助优化该任务的其他信息。对于不同的任务来说，有些任务之间是存在相关性的，例如在语音识别中，文献[60]中的研究表明人的情绪状态被认为是性别依赖的，即说话人的情感和说话人的性别必然有着内部的关联性，很难把两个识别任务

完全地分离开。因此在进行单独任务的识别时，必然会忽略某些语音中的共有特征，造成识别模型存在一定的局限性。

3.3.2 多任务学习

如上节所述，现实中很多问题之间存在着或多或少的联系，当我们把它们分隔开来看待时，往往会忽略问题之间所存在的关联信息。把解决问题作为任务，那么存在相关性的任务就叫做相关任务。为了解决上述问题，参考人类感知和学习模式的特点，研究者们提出了多任务学习的概念。多任务学习是迁移学习（Transfer Learning）算法的一种，本质上是一种归纳迁移机制，利用隐含在多个相关任务中的特定信息来提高泛化能力[58]。在神经网络中，往往通过使用共享表示并行训练多个相关任务来实现多任务学习。额外任务的信息有助于网络学到更好的内部表示，使得神经网络能够通过共享隐层学习到更多共享特征，来提升各个任务的性能，提高泛化能力。同时，由于多任务学习在相关任务间共享网络参数，在对多个任务进行预测或分类时，训练所需的数据量和模型参数的数量都将明显减少，使模型更加高效。对于上节所述的说话人性别、情感、身份三个识别任务的场景，采用多任务学习结构的模型实现方式如图 3.5 所示，将数据集输入多任务神经网络模型，三个识别任务在共享隐层中共享网络参数，然后在输出层中分别输出各任务的分类结果。

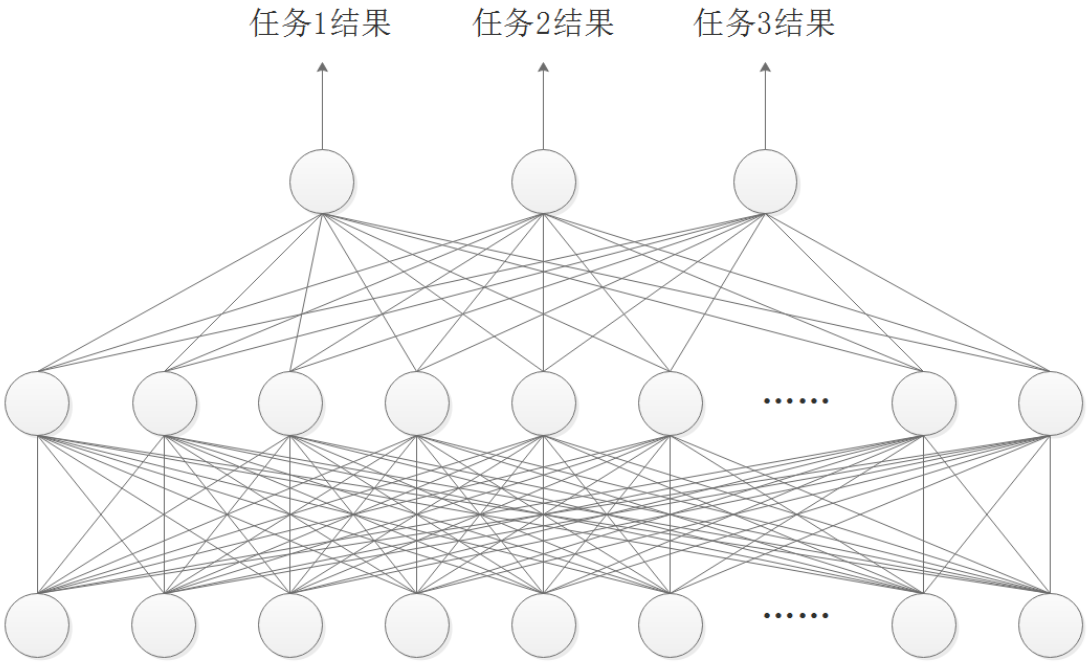

图 3.5 多任务学习结构

相对于图 3.4 中的单任务识别模型，图 3.5 中的多任务学习结构有几大优势：第一，多任务学习模型可以学习到相关任务之间的共有特征，模型泛化能力更强，仿真实验可以证明各

任务的性能都有所提升；第二，单任务识别模型的多个模型需要重复的对语音信号数据进行处理、特征提取、训练等操作，多任务识别模型可以并行处理多个任务，只需要一次共同的数据特征处理和训练操作，提高了模型的识别效率，缩短了模型运行时间；第三，神经网络训练过程中经常存在的另一个问题是模型过拟合，而多任务学习模型同时学习和表征多项任务的特征，减小了模型的过拟合程度[72]。

由于上述优点，多任务学习模型被广泛的应用于人工智能领域[28,32,52,75]。多任务学习有多种形式，例如联合学习（Joint Learning）、自主学习（Learning to Learn）和带有辅助任务的学习（Learning with Auxiliary Task）等，其本质都为同时优化多个损失函数，只是在总的损失函数的构造上不同。目前，多任务学习研究的难点在于寻找同时适合多个识别任务的特征参数以及如何更好的挖掘和利用这些共享特征。

3.3.3 含有属性依赖层的多任务神经网络模型

在常规多任务学习的基础上，针对不同的目的，研究者们提出了各类多任务学习结构。其中，文献[28]提出了一种含有属性依赖层的多任务学习，如图 3.6 所示。以两层隐层的简单多任务网络举例，左图是常规结构的 MTL，两层隐层都为共享层，三个任务共享其中的网络参数，右图是属性依赖的 MTL，第一层隐层作为共享层，而第二层隐层分为三个独立的隐层，文献中称之为属性依赖层，分别供三个任务独立训练，目的是使网络中的每个任务在学习到共享特征的同时，允许每个子任务在属性依赖层独自优化并学习到独有特征以提升性能。文献中的实验结果表明，采取了属性依赖结构的多任务神经网络模型，在其他条件相同的情况下，相较于普通的多任务神经网络各任务均有更好的性能。

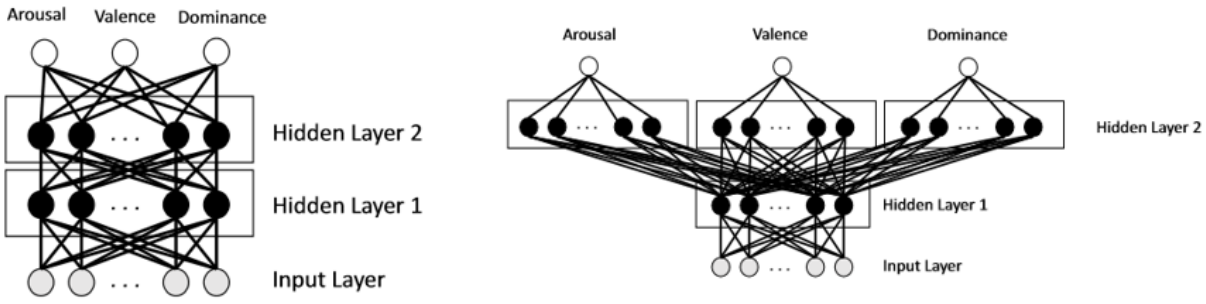

图 3.6 含有属性依赖层的多任务学习[28]

3.3.4 本文多任务神经网络模型构建

在现有的将多任务学习运用于语音识别领域的研究中，往往以选取某些辅助任务来提升

主任务性能的方式进行单维语音识别，例如采用情感属性的等级划分作为辅助任务来提升主任务情感识别的性能；且一般采用全连接网络或卷积神经网络进行特征提取和分类。本团队成员在文献[3]已经进行过通过多任务学习理论实现多维语音识别的研究，其采用了常规的多任务学习结构来构建模型，但其网络结构较为简单，为仅有两个共享隐层的 DBN 模型，在模型结构以及性能上有改进和提升的空间。

综合上文的研究和分析，本文提出了一种新的用于多维语音识别的多任务神经网络模型，采取性能更好的带有属性依赖层的多任务学习结构。其中，采用适合处理时序数据、表征语音信号特征的 RNN 来构建共享层，使各相关识别任务能够有效学习共有特征，并选用 GRU 作为 RNN 的基本单元；采用全连接网络构建属性依赖层，使各识别任务能够独自学习私有特征。最终模型同时输出多个语音识别任务的识别结果。

3.4 基于多任务学习和循环神经网络的多维语音识别模型

3.4.1 算法模型设计

本章提出了一种基于多任务学习和循环神经网络的多维语音识别算法（MTL-RNN），采用了适合处理连续语音数据的循环神经网络和具有属性依赖层的多任务学习结构。模型所选择的语音识别任务是说话人性别识别、说话人身份识别、说话人情感识别。关于识别任务的选取，首先人的情绪状态被认为是性别依赖的，已有研究表明特定于某种性别的情感分类器的性能明显优于不分性别的情感分类器[60]；其次，不同说话人对相同情感的表达必然存在着共同和不同的部分，对说话人身份的确定有助于理解不同情感中不同的部分，对某种情感的表达方式也有助于确认说话人的身份。因此这三项语音识别任务互为相关任务，满足多任务学习的根本原则。故本文选择针对这三项语音识别任务构建基于多任务神经网络的多维语音识别算法模型。对于其中的说话人身份识别来说，本文所研究的类别属于文本无关的说话人辨认（即说话人可以说任意文本），以最大限度地满足实际应用场景；对于说话人情感识别来说，本文在情感的分类方式上选取的是离散情感模型，这样在实验所用语音库的选取方面有更多的自由性。

本团队成员在先前的研究中已经提出过一些多维语音识别模型。文献[2]中采用的多示例多标记支持向量机算法构建多维语音识别模型，先进行说话人性别识别，再进行基于性别的说话人情感和身份的同时识别；文献[3]中构建了两个多维语音识别模型，一个是通过多任务学习结合 DBN 构建了两层共享隐层的多维语音识别模型进行说话人性别、身份、情感的同时

识别；另一个是基于 ProgNets 结构建立模型进行基于性别分类的说话人情感和身份的同时识别。与这些模型相比，本章提出的 MTL-RNN 模型的创新性在于将适合处理时序数据的循环神经网络和有效提升各任务性能的带有属性依赖层的多任务学习结构相结合，采用 RNN 作为网络共享层，选择 GRU 作为 RNN 基本单元，采用全连接网络作为属性依赖层，充分利用各识别任务之间的相关性，进行说话人性别、身份、情感的同时识别。

本章研究所构建的基于多任务学习和循环神经网络的多维语音识别模型的整体系统流程框图如图 3.7 所示。模型整体采用带有属性依赖层的多任务学习结构，使得网络模型在 RNN 共享层中学习多个任务的共享特征的同时，每个子任务在全连接属性依赖层能够独自学习私有特征、优化参数以提升分类性能。首先，将输入语音进行预处理以及提取特征。在特征选择方面，经过对单维语音识别的研究，目前被广泛使用的 MFCC 特征在说话人性别识别、身份识别、情感识别三个种类的识别任务中均有不错的表现，因此在本章的研究中采用组合 MFCC 参数作为多维语音识别模型所用的语音特征。

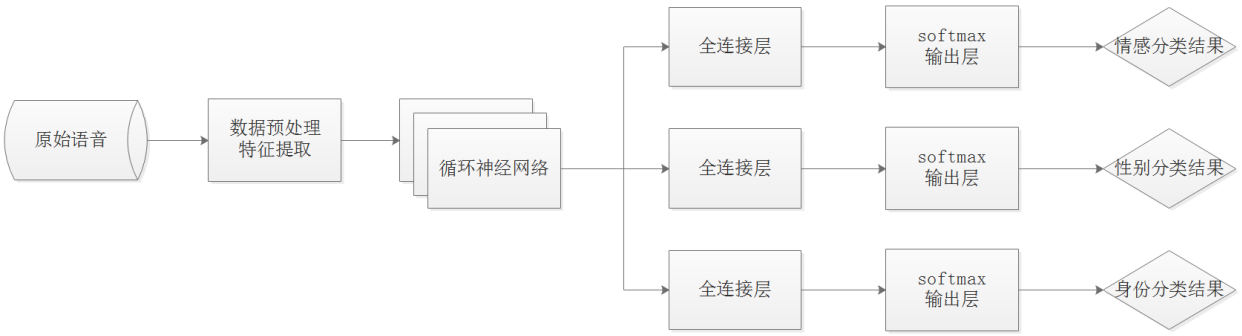

图 3.7 基于多任务学习和循环神经网络的多维语音识别模型流程图

将 WAV 格式的音频文件输入模型中后，首先使用语音活动检测器消除实验语音中多余的静音部分；随后对语音信号以 16kHz 重采样后再进行分帧，帧长为 512 个采样点，帧叠为 256 个点；接着将每一帧乘以汉明窗，并进行快速傅里叶变换；接下来通过 mel 滤波器组、计算对数能量、经过离散余弦变换（Discrete Cosine Transform, DCT）后得到 13 阶的梅尔倒谱系数，然后计算 MFCC 特征的一阶差分和二阶差分；最后将 13 维的 MFCC、13 维的一阶差分、13 维的二阶差分以及语音帧能量作为一帧的 40 维组合 MFCC 特征。随后将得到的组合 MFCC 特征向量按照时间顺序以每帧为一个时间步输入到共享层中，本章使用三层循环神经网络作为共享层，采用 GRU 作为基本单元，每层 512 个节点。其中，第一层和第二层输出全部时间上的序列结果到下一层，第三层输出最后一个时间步的结果并将其输入到属性依赖层中。属性依赖层由三个单独的全连接层组成，各有 512 个节点，分别将结果输出到三个单独的 softmax 输出层中，网络的最后同时输出三个任务的分类结果。

在训练阶段，将语音训练集数据以批次为单位送入神经网络中训练，将输出层输出的分类结果与训练语句的分类标签做比较并计算损失，随后使用所选取的优化器对损失函数进行优化，更新网络权重，更新完毕后进行下一批次的训练，直到模型收敛，得到最小的损失值。在测试阶段，以同样的方式对测试集提取特征然后输入到网络中，得到输出结果后与测试语句的分类标签进行比较，并输出三类识别任务的测试结果。

3.4.2 损失函数定义

多任务学习从实质上来说就是网络模型同时优化多个识别任务的损失函数（Loss Function），将每个任务的损失函数融合成一个总的损失函数，使用优化器基于误差反向传播的基本理念对损失函数进行优化，以此来更新网络权重参数。本章针对每一个识别任务采用交叉熵函数作为损失函数，总的损失函数可以表示为。

$$loss_{total} = \alpha loss_{emo} + \beta loss_{gen} + \gamma loss_{spe} \tag{3.11}$$

其中 $loss_{emo}$、$loss_{gen}$、$loss_{spe}$ 分别为语音情感识别、性别识别、身份识别的损失函数，由下式计算得出：

$$loss = -\sum_{i=1}^{P} y_i \log(\hat{y}_i) \tag{3.12}$$

其中 y_i 目标值，$\hat{y}_i$ 为神经网络输出的预测值，P 为该任务的类别总数。α、β、γ 为权重系数，三者的和为 1。权重系数的取值在仿真实验中通过验证集进行调试和确定，根据不同语音库的特点和不同的目的，可以对各识别任务的权重进行灵活调整，使得模型收敛时的总损失最小。

3.5 实验结果分析

3.5.1 实验语音数据库

多维语音识别在实验中所用的语音数据库相较于单维语音识别有更高的要求，所用的语音库必须具备多种说话人信息，具有多种语音识别任务的分类标签，能够满足多个不同语音识别任务的训练和测试要求。由于目前在多维语音识别领域的研究较少，也暂未出现专门为多维语音识别研究而录制的专门的多维语音库，因此需要在语音识别领域常用的语音库中寻找适合进行多维语音识别研究的语音数据库。经过研究发现，将说话人进行区分的情感语音

库可以满足本文研究中对说话人身份识别、情感识别、性别识别的多维语音识别的要求。该领域常用的经典语音库有柏林语音情感数据库（Emo-DB）等，可惜的是其语音样本数据量较少，而神经网络模型因为节点和连接多需要较大规模的样本数据来对模型进行充分的训练，因此 Emo-DB 并不适合本文的研究。

经过调查和筛选，本文的仿真实验首先使用了中国科学院自动化所录制的汉语情感语料库（CASIA）。该语料库由四名说话人（两男两女）录制而成，语音数据包括六种情感，分别为生气（anger）、害怕（fear）、开心（happy）、中性（neutral）、悲伤（sad）、惊讶（surprise）。每个说话人每种情感有 300 条相同文本的语句，共 7200 条语句，全部数据均用于了实验。仿真中将实验数据按照 6:2:2 的比例分为训练集、验证集和测试集。训练集用来对模型进行训练，验证集用来确定模型中的各项超参数，测试集用来评估模型性能。

虽然 CASIA 语音库的语音数据量较大，但存在说话人数量较少的问题，对本文多维语音识别研究中的说话人身份识别任务来说难度较小，难以发现其中的问题，因此，为使本章模型性能更具有说服力，另一方面也为了体现本章的多维语音识别模型在不同语种的语音数据上的性能表现，本文还在现代标准阿拉伯语（Modern Standard Arabic）的沙特阿拉伯国王大学情感语音库（King Saud University Emotions，KSU）[61]上进行了仿真实验。实验使用了该语料库中的第二组语音数据，其由 14 个说话人（7 男 7 女）录制而成，包括五种情感，分别为生气（anger）、开心（happy）、中性（neutral）、悲伤（sad）、惊讶（surprise）。每个说话人每种情感有 24 条语句，共 1680 条，全部数据均用于了实验。同 CASIA 语音库一样，本章在仿真实验中同样将 KSU 语音库数据按 6:2:2 分为训练集、验证集和测试集。

3.5.2 实验环境参数设置

本章采用基于 python 实现的 librosa 语音信号处理工具对语音数据进行预处理和组合 MFCC 特征的提取，其中提取的组合 MFCC 特征的维度为 40 维。神经网络模型在训练阶段采用小批次 ADAM（mini-batch ADAM）优化算法[62]对损失函数进行优化，批次（batch）大小设置为 50；每 50 个批次测试一次每项任务分别在训练集和测试集上的准确率；网络除输出 softmax 层采用 softmax 函数作为激活函数以外，其余层均采用 tanh 函数作为激活函数；设定基础学习率为 0.001，学习率按照 0.99 的系数每 50 个批次衰减一次；为减轻网络的过拟合问题，网络模型除输出层以外均采用了 0.5 的 dropout 结构。实验在 python 环境下进行，使用 python3.6 中的 GPU 版 TensorFlow 框架进行多维识别模型的训练和测试，进行多次实验取平均值作为实验结果。

对于上文所述的多维识别模型总损失函数中权重系数的取值，本章通过验证集的数据对其进行确定。对于CASIA语音数据库的数据，本章在实验中首先取α、β、γ的值为0.5、0.25、0.25作为权重初始值，对网络进行训练，并在验证集上测试网络性能，随后根据性能表现调整权重值并继续在验证集上测试，最终找到使模型在验证集上取得最高识别率时相对应的α、β、γ的值，经过实验后得到α、β、γ的最终取值为0.5、0.15、0.35。对于KSU语音数据库来说，说话人数量比CASIA多出很多，情感类别减少为五个，因此可以通过调整权重的取值来使模型更加适合KSU语音库。经过相同的验证集操作后，在KSU语音库上的实验中α、β、γ的最终取值为0.25、0.1、0.65。

3.5.3 对比仿真实验设置

为体现本章提出的基于多任务学习和循环神经网络的多维语音识别算法（MTL-RNN）的性能，除对本章所提出的模型进行实验外，分别设计了五组对比仿真实验，其中均使用提取方法相同的40维组合MFCC特征：

（1）为验证基于多任务学习实现的多维语音识别相较于单维语音识别在性能上的提升，设计单维语音识别对比实验STL-RNN，在模型结构和其他超参数保持相同的前提下，分别对说话人情感、性别、身份三项识别任务进行单任务识别，并记录各项识别任务的识别率。

（2）为凸显RNN对于时间连续数据的处理能力，设计全连接网络对比实验FCN，使用结构、层数、节点数与本文模型相同的全连接多任务神经网络对情感、性别、身份三项分类任务进行分类。

（3）为体现多任务结构中属性依赖层的有效性，设计对比实验Only-RNN，在保证其他超参数相同的条件下，使用RNN构建没有属性依赖层的常规多任务神经网络进行实验，记录实验结果。

（4）为表现属性依赖层和共享层在功能上的不同，设计对比实验Share-RNN，将原算法中的属性依赖层更改为全连接共享层，由该层输出到 softmax 输出层得到三个任务的分类结果，并保持其他超参数不变，对比实验结果。

（5）为测试本章提出的MTL-RNN模型在噪声环境下的性能表现，设计几组噪声对比实验，通过将实验用的KSU语音库中的测试组数据人为加入不同信噪比的高斯白噪声，在使用无噪声数据对模型训练完毕后，将测试的混合声音信号输入到模型中进行多维识别，其中设置的信噪比从0dB到35dB共8组实验，记录并对比实验结果。

3.5.4 结果分析

（1）多维语音识别模型和单维语音识别模型的性能比较

首先将本章提出的多维语音识别模型和相同条件下的单维语音识别模型性能进行比较。表 3.1 分别显示了本章模型和单维语音识别模型在 CASIA 和 KSU 两个数据库上各项识别任务的识别率（本章和第四章的实验结果表格中的“平均”一栏表示三项识别任务的平均识别率）。

表 3.1 单维模型和多维模型在 CASIA 和 KSU 上的各项任务识别率

语音库	模型	性别	身份	情感	平均
CASIA	MTL-RNN	99.99	99.09	90.87	96.65
	STL-RNN	99.36	98.06	83.51	93.64
KSU	MTL-RNN	99.70	95.83	90.48	95.34
	STL-RNN	99.05	89.78	81.91	90.25

由表 3.1 可以看出，本章提出的基于多任务神经网络的多维语音识别模型 MTL-RNN，在两个语音库上，三项识别任务的识别率相对于单维识别均有不同程度的性能提升。在 CASIA 语音库上进行的实验中，多维模型的平均识别率为 96.65%，比单维识别高出 3.01%；在性别识别上，由于该识别任务难度较低，多维模型和单维模型的识别率均接近百分之百，但多维模型仍有 0.63%的性能提升；在身份识别上，多维识别模型的识别率相对单维识别提升了 1.03%，由于 CASIA 说话人数量较少，所以多维模型和单维模型的身份识别率均较高；在情感识别上，多维模型的识别率高出单维模型 7.36%。在 KSU 语音库上进行的实验中，多维模型的平均识别率为 95.34%，比单维识别高出 5.09%；在性别识别上，多维模型和单维模型识别率同样接近百分之百，多维模型的识别率高出了 0.65%；在身份识别上，由于 KSU 语音库中说话人数量的大幅增加，多维识别模型的身份识别率相对于 CASIA 库略有所降低，但相对于单维识别的识别率提升也更多，为 6.05%；在情感识别上，多维模型的识别率比单维模型高出 8.57%。总而言之，多维语音识别模型充分利用了相关任务之间的相关信息，在不同语种的语音库中各项识别任务均取得了高于单维语音识别模型的识别率，证明了多维语音识别算法模型的有效性。

（2）带有属性依赖层的多任务神经网络相关性能分析

为了对含有属性依赖层的多任务神经网络进行有效性分析，本章设置了三个对比实验，即全连接网络 FCN、没有属性依赖层的常规多任务学习网络 Only-RNN 以及共享层代替属性依赖层的 Share-RNN。实验结果如表 3.2 所示。

表 3.2 带有属性依赖层的多任务神经网络相关实验结果

语音库	模型	性别	身份	情感	平均
CASIA	MTL-RNN	99.99	99.09	90.87	96.65
	FCN	80.33	76.27	66.28	74.29
	Only-RNN	99.89	98.51	87.15	95.18
	Share-RNN	99.99	98.71	87.37	95.36
KSU	MTL-RNN	99.70	95.83	90.48	95.34
	FCN	69.31	57.63	68.11	65.01
	Only-RNN	99.14	93.12	88.13	93.46
	Share-RNN	99.08	93.16	87.98	93.41

从表中数据可以看出，本章提出的 MTL-RNN 模型在两个语音数据库上各项识别任务相对于各对比实验均有不同程度的性能提升。其中，由于单纯的多任务全连接网络 FCN 并未有任何特殊结构或其他分类器，而是仅将提取的 MFCC 特征用全连接网络进行分类，故其识别率较低；基于 RNN 的几个网络均有较高的识别率，说明 RNN 对于连续语音数据的描述能力明显强于普通的全连接网络；Share-RNN 模型和 Only-RNN 模型在识别率上差别较小，说明在训练数据有限的情况下，在 RNN 共享层之后加上一层全连接共享层对于本实验语音库上的模型性能影响较小；MTL-RNN 模型对于 Only-RNN 和 Share-RNN 模型的识别率提升说明了 RNN 共享层之后的属性依赖层强化了模型对每项任务的学习，提升了模型的性能。另外，由于 KSU 库中说话人数量的大幅增加，MTL-RNN 模型对于性别和身份的识别率相较于各对比实验的提升比 CASIA 库中更加明显。

（3）多维识别模型在噪声环境下的性能表现

本章所进行的语音识别研究和目前大多数研究一样，采用了公开的语音数据库进行仿真实验，而这些语音库往往是在环境背景噪声较为干净的环境下录制。为了测试本章提出的模型在噪声环境下的性能表现，即系统的鲁棒性测试，通过向 KSU 数据库实验中的测试集数据中人为添加指定信噪比的高斯白噪声，训练集和验证集不做任何改变，测试模型各项任务的识别性能。实验结果如表 3.3 所示。

表 3.3 多维识别模型在不同信噪比的噪声环境下的性能

信噪比	性别	身份	情感	平均
0dB	77.34	22.31	44.74	48.13
5dB	88.26	29.65	60.70	59.54
10dB	90.14	42.23	64.68	65.68
15dB	96.12	52.07	67.33	71.84
20dB	98.34	61.61	73.34	77.76
25dB	98.65	72.54	76.37	82.52
30dB	98.81	90.44	82.12	90.46
35dB	99.41	95.73	89.01	94.72
无噪声	99.70	95.83	90.48	95.34

由表中数据可以看出，随着信噪比的增加，模型的平均识别率总体处于不断上升的状态；当信噪比在 5dB 以下时，模型性能的降低较多，平均识别率不足 60%；当信噪比高于 30dB 时，模型的平均识别率逐渐接近无噪声状态下的平均识别率值。总体来说，本章提出的多维识别模型具有一定的抗噪性能，但还可以在抗噪声性能方面进行改进完善，这也是后续研究的方向。

（4）同类研究性能比较

为了体现本章模型的性能优越性，将本文模型的仿真结果与本团队之前的同类型研究进行性能比较，与文献[3]中同样使用 KSU 语音库进行说话人情感、性别、身份三项识别任务的多维识别模型进行各项识别任务的识别准确率比较，结果如表 3.4 所示。

表 3.4 同类研究性能比较

模型	性别	身份	情感	平均
本文模型	99.70	95.83	90.48	95.34
文献[3]	98.56	92.85	85.71	92.37

由表中数据可以清楚的看出，本章提出的多维语音识别模型的平均识别率比文献[3]高出了 2.97%，在各项任务上均有着更好的识别准确率，证明了本章提出模型的优越性。

3.6 本章小结

本章提出了一种同时对说话人身份、情感、性别三种说话人特征信息同时进行识别的多维语音识别方法（MTL-RNN），将多任务学习机制与循环神经网络结构相结合，充分利用了

声音信号中丰富的信息，有效提升了各项任务的识别性能。本章首先对所提出模型中涉及到的多任务学习和循环神经网络的基本概念和理论进行了介绍。接着介绍了本章提出的基于多任务学习和循环神经网络的多维语音识别模型。语音信号数据经过预处理后提取 MFCC 特征并输入到模型中，通过 RNN 共享层共享网络超参数以学习各任务的共享特征；通过全连接属性依赖层学习各任务独有特征；通过调整模型总损失函数中各任务损失函数的权重来对不同特点的语音数据库进行针对的优化；最终同时输出说话人身份、性别、情感三项识别任务的分类结果。最后，对本章提出的 MTL-RNN 模型和设置的对比实验进行仿真实验，实验结果表明，本章提出的基于多任务学习和循环神经网络的多维语音识别模型的平均识别率在两个语音库上分别比单维识别模型高出 3.01%和 5.09%，三项识别任务均有不同程度的识别率提升，对于语种因素和说话人的个性因素有较好的鲁棒性，且具有一定的抗噪性能。

第四章 基于卷积神经网络和特征融合的多维语音识别

本章从识别模型所用的语音特征参数改进入手，继续通过多任务学习机制和循环神经网络构建模型，另外通过卷积神经网络对语音信号的语谱图进行特征提取，将提取的特征向量与 MFCC 特征进行融合，得到更适合进行多维识别的特征，最后将融合特征输入到多任务神经网络中进行多维语音识别。

4.1 引言

在本文第三章中，运用多任务学习机制，采用带有属性依赖层的多任务网络结构，结合循环神经网络，构建了多维语音识别模型，实现说话人身份、性别、情感三种识别任务的同时识别，最终在不同语种的语音库上均取得了优于单维识别的结果，验证了多维语音识别的可行性。如第二章中所述，在语音识别的研究中，最为关键的两个部分是语音特征提取和分类模型构建，第三章的主要研究点在于构建可以充分利用相关任务之间的相关信息的多维分类模型，在语音识别特征选择方面选取了在语音识别领域使用较为广泛的 MFCC 组合特征作为特征参数。然而多维语音识别所用的语音特征则需要尽可能地包含多维度的语音信息，尽可能同时满足各项识别任务的分类要求，以便多维模型能够学习到语音信号中各识别任务的共有特征和独有特征，更好地完成多维语音识别任务。显而易见，在多维语音识别的研究中，简单的采用 MFCC 特征作为特征参数或者如同类研究文献[3]中采用 i-vector 特征的方式是有所欠缺的。因此，本章中主要着手进行多维语音识别特征参数改进的研究。

关于语音识别常用的特征参数，本文在第二章中已经进行了详细的介绍。考虑到多维语音识别特征参数的要求，对语音识别领域常用特征进行分析，注意到谱特征中包含丰富的语音信息，而常用的谱特征如 MFCC、Mel-Filterbank、PLP（Perceptual Linear Prediction）等的提取流程中有共同的处理步骤，即对语音信号进行预处理之后进行傅里叶变换，之后再进行不同的滤波等处理。而语音信号经过短时傅里叶变换后得到的是语音信号的语谱图。语谱图表示了语音信号的频谱随时间的变化，相对于上述经过各类手段提取的谱特征，语谱图包含了更多更丰富的原始语音信息，更符合多维识别特征参数的要求。然而另一方面，语谱图中的信息也可能存在冗余的问题。这时，自然联想到神经网络中适合对特征进行筛选、降维、提取的卷积神经网络。

随着计算机计算能力的提升，卷积神经网络开始在图像识别领域大放异彩，完全代替了传统的图像识别方法，可以直接对输入的图像进行特征提取和降维，在处理方式上非常拟人

化，并取得了非常好的效果。而在语音识别领域中，卷积神经网络同样可以有很好的应用。可以将上文提到的语音信号语谱图当作图像输入到卷积神经网络中处理，进行特征提取和降维，这样既能保留语谱图中丰富的多维语音信息，又能避免特征冗余的问题。

因此，本章研究的工作主要在多维语音识别模型的特征参数提取的改进上，先使用卷积神经网络对语音信号的语谱图进行新型多维特征提取，并输入到第三章中的多维语音识别模型中进行说话人性别、情感、身份三项识别任务的分类。接着分析了单一特征参数的不足，介绍了特征融合的方法，并将基于卷积神经网络得到的特征向量与 MFCC 特征进行融合，输入到多维语音识别模型中进行识别，同样得到三类识别任务的分类结果。最后对本章提出的多维语音识别方法进行仿真实验，设置多个对比实验，比较实验结果，显示本章提出的特征提取和融合方法的有效性和先进性。实验表明，本章提出的基于卷积神经网络和特征融合的多维语音识别模型，对比第三章中的采用 MFCC 作为特征参数的多维识别模型，在三项识别任务上均有一定程度的性能提升，体现了本章提出的融合特征用于多维语音识别的有效性。

4.2 卷积神经网路

4.2.1 卷积神经网络概念

卷积神经网络 CNN 本质上是一种包含卷积计算并具有深度结构的前馈神经网络，是深度学习的代表算法之一[62]。CNN 具有表征学习（Representation Learning）的能力，能够按其阶层结构对输入信息进行平移不变分类（Shift-invariant Classification），因此也被称为“平移不变人工神经网络（Shift-Invariant Artificial Neural Networks, SIANN）”。

最早的 CNN 相关研究可以追溯到 1980 年左右，经过一段时间的研究和探索，研究者 LeCun 于 1989 年正式使用“卷积”这一说法来描述网络结构[63]。随后，LeCun 等人于 1998 年提出的 LeNet-5 的 CNN 结构定义了现代卷积神经网络的基本结构[64]。近年来，随着计算机性能的飞跃，卷积神经网络在图像识别等领域取得巨大成功，各类先进、复杂的 CNN 结构被不断的提出和改进，例如 AlexNet[65]、VGGnet[66]、ResNet[67]等。

相对于传统的方法，CNN 的优势在于其不需要人为进行特征的筛选和提取，可直接将图像或语音以某种形式输入到网络中，由网络通过训练对特征进行自动提取和处理。CNN 有以下几个特点：第一是局部连接，即每个神经元不和上一层的所有神经元相连，而是只和其中一部分神经元相连；第二是参数共享，在进行特征的卷积操作时，同一层的所有神经元可以共享一组卷积核参数；第三是下采样，通过池化操作减少特征数量，保留有用信息的同时剔

除冗余信息。由于以上特点，CNN 在拥有强大特征表征能力的同时，能够有效减少网络模型参数数量，提升模型效率。

在 2012 至 2013 年左右，研究者们开始将 CNN 引入到语音识别领域中，用来对语音特征进行提取、降维、筛选、学习和处理，在实际使用中往往通过 CNN 结合其他类别神经网络如 DNN、LSTM 等来构建语音识别分类模型，从而获得更好的模型性能，其中比较经典的应用有 CLDNN[54]、DeepSpeech[97]等。

本章同样采用多种神经网络分类器组合的方式来构建多维语音识别模型，采用 CNN 进行特征参数提取并结合 RNN、全连接网络进行多任务分类。与上述模型相比，本章模型的创新之处在于：第一，本章模型中的 CNN 根据实验数据集和分类任务的特点设计了对应的层数、每一层的参数等，其结构与上述模型不同；第二，本章模型采用了多任务学习结构实现多维语音识别，上述模型往往用于内容识别等单维识别；第三，本章模型将 CNN 提取的特征与组合 MFCC 特征进行特征参数融合，再输入后续的 GRU 网络中进行分类，因此提取的特征参数不同。

4.2.2 卷积神经网络结构

CNN 的基本操作包括卷积（Convolution）、池化（Pooling）、激活等，通常情况下，一个 CNN 模型包含多层的卷积层和池化层，经过多次激活函数（Activation Function），最后一层一般为分类器。一个典型的 CNN 模型流程结构如图 4.1 所示。

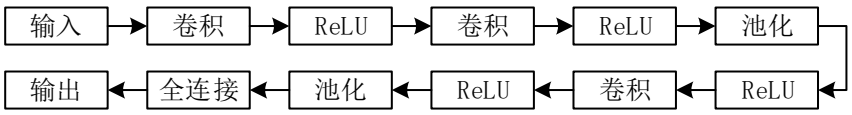

图 4.1 典型 CNN 结构

CNN 通过多层卷积抽取模型输入特征向量中的复杂特征，通过多次池化对特征向量进行降维，从而完成对输入数据的处理并进行高效的特征提取。与常规神经网络相同，CNN 也采用反向传播算法，通过梯度下降法优化损失函数值来对模型中的各项权重参数进行更新。上图流程中每个步骤的操作如下。

（A） 输入

CNN 可以处理多维度的数据，其输入可以是一维语音采样、某种语音特征参数，也可以是二维图像、语谱图等，三维的 RGB 图像和四维的数据同样也可以处理。由于 CNN 使用梯

度下降法对网络权重参数进行优化，因此需要对输入特征数据进行归一化处理，以便网络模型能够更好的发挥学习能力。

（B） 卷积

卷积层和池化层是 CNN 区别于其他类型神经网络的、独有的结构，其中卷积层更是 CNN 的核心。卷积操作的实质就是通过多个卷积核（Convolutional Kernel）或者说卷积滤波器（Convolutional Filter）与输入的特征向量进行卷积计算。图 4.2 为二维卷积的示意图。

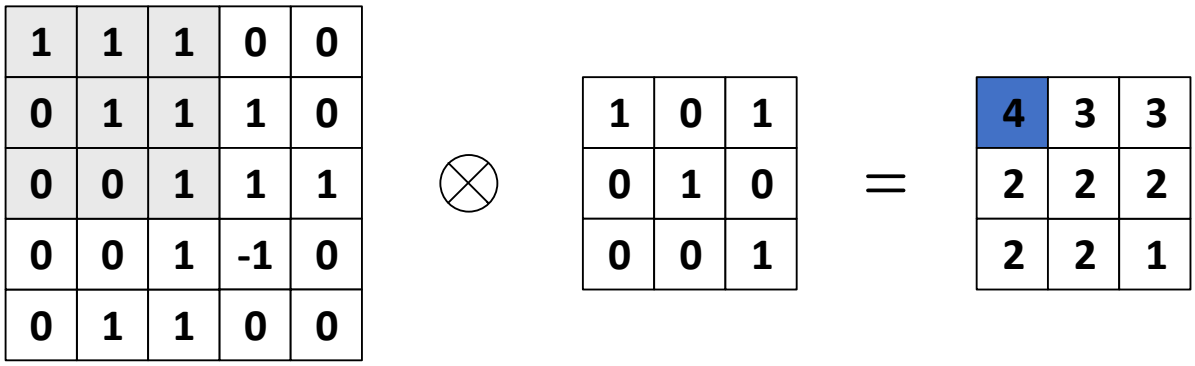

图 4.2 二维卷积示例

左边为输入的二维特征，中间为一个 3×3 的卷积核，卷积操作即是卷积核中的值和特征向量对应位置的值（图中阴影部分）计算点积，得到右侧第一行第一列的卷积结果值。随后，卷积核向右移动，移动的长度叫做步长，再将对应位置的值计算点积，得到第一行第二列的值。以此类推，多次计算后得到的新的特征即为该卷积核最终的卷积结果。在实际应用中，往往设置一组多个卷积核，可以认为不同的卷积核提取的是不同的特征。另外，卷积操作还有一个概念叫做填充（Padding）。可以发现，卷积核的尺寸和步长决定了输出特征的维度，经过卷积后输出的特征尺寸可能会减小，如果这不是所期望的，则可以通过填充操作在输入特征的边界以外补充零值（零填充，如左侧第一张图阴影部分），以达到输出特征尺寸和输入保持一致（或控制输出特征尺寸，如最右侧图阴影部分）的目的，如图 4.3 所示。

卷积操作可以提取出输入参数的特征，多个卷积层相叠加可以提取输入参数的复杂特征，这是 CNN 有效性和先进性的核心所在。

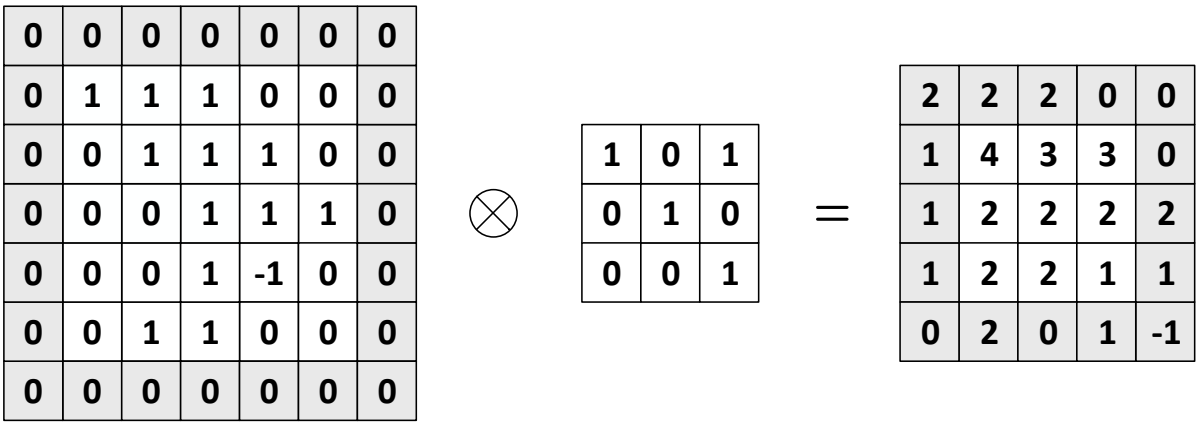

图 4.3 填充操作

（C） 激活函数

在卷积层之后，往往会通过一个激活函数来协助模型更好的表达特征。这里的激活函数与其他类型神经网络中的激活函数概念相同，其作用是加入非线性因素。在 CNN 中通常使用线性整流函数（Rectified Linear Unit, ReLU）作为激活函数，其表达式如下式所示。

$$f(x)=\begin{cases}x, x\geq 0\\ 0, x<0\end{cases} \tag{4.1}$$

（D） 池化

在若干个卷积层后，往往会添加池化层，其目的是对特征进行降维、减少网络权重数目、减少计算复杂度以及防止过拟合，有助于 CNN 提取深层次的特征。池化的种类包括最大池化（Max-Pooling）、平均池化（Mean-Pooling）等，图 4.4 为 2×2 最大池化的示意图，即对每块指定区域大小的特征数据，保留其中的最大值（第一步池化左侧阴影区域最大元素值是 3，对应池化后右侧图中第一个阴影位置的 3；以此类推），作为池化后的输出。通过池化操作，可以很大程度上减少模型的计算量和特征的尺寸，保留有用信息，剔除冗余信息，且使模型的泛化能力更强。

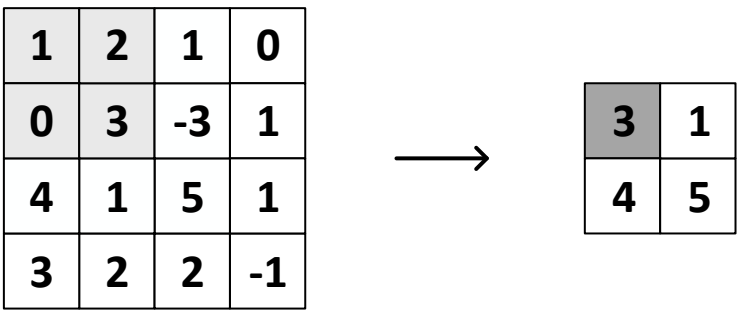

图 4.4 最大池化操作

（E） 全连接网络

图中最右侧的全连接网络在模型中起到分类器的作用，利用前面通过卷积和池化得到的高阶特征进行分类，得到模型分类结果。如果有额外设置的其他类型分类器，全连接网络还可以起到将输出的二维局部特征映射为一维全局特征的作用。在本章的研究中，采用了 RNN 和全连接网络相结合来构建分类器。

4.3 基于卷积神经网络的端到端多维语音识别模型

4.3.1 算法模型设计

本节提出了一种基于卷积神经网络的端到端多维语音识别算法，其中端到端指的是算法模型输入是语音信号（实际操作时输入是语音的原始语谱图），输出是语音识别任务的分类结果，中间的所有步骤都包含在神经网络内部进行处理，不需要额外的人工特征选择等操作。本算法模型在特征提取上采用了 CNN 对语音信号语谱图进行特征提取、筛选和降维，并将经过 CNN 提取的语音特征参数输入与第三章中的模型类似的 RNN 多任务神经网络分类器中，完成对说话人性别、身份、情感这三项识别任务的分类。

在第三章的研究中，使用了 MFCC 特征作为多维语音模型的特征参数，取得了一定的效果，证明 MFCC 对三种任务都具有一定的区分性。然而 MFCC 是通过人工提取的语音特征，使用了许多信号变换的手段，摒弃了许多语音信号中丰富的语音信息，这与多维语音识别理念不完全吻合，多维语音识别期望尽可能的利用到语音信号中的多维度信息来进行拟人化的识别。考虑到上述问题，本节的研究一方面采用了包含丰富语音信号原始信息的语音语谱图；另一方面，由于 CNN 具有卓越的图像分析处理能力，本节采用 CNN 对语音信号语谱图进行特征提取，得到多维识别模型所用的多维特征。最后将 CNN 的输出特征输入到后续的分类器中进行多任务分类。

4.3.2 算法流程和参数设置

本节研究构建的基于卷积神经网络的端到端多维语音识别模型的整体系统流程图如图 4.5 所示。将原始语音信号的 WAV 格式的音频文件输入到模型中，进行语音活动检测消除静音部分，接着同第三章一样进行 16kHz 重采样，以 512 个采样点分帧，帧叠为 256 个采样点，通过短时傅里叶变换得到语音信号的语谱图，这里设定语谱图的尺寸为 $N\times128$，N 为帧数，将其输入到 CNN 中。

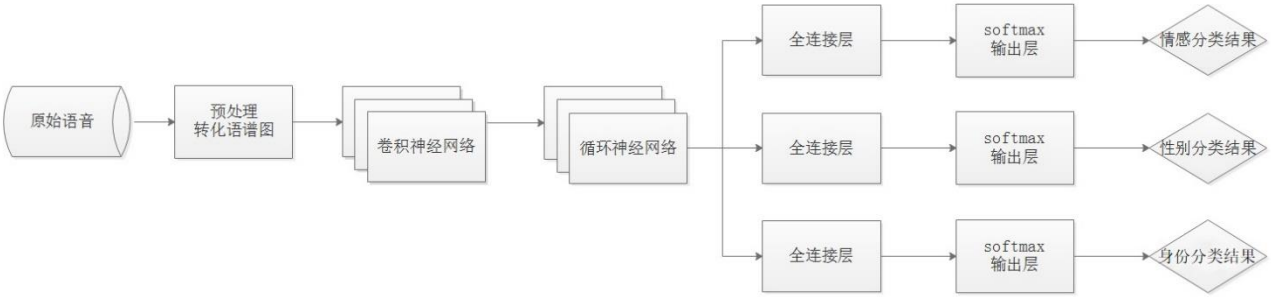

图 4.5 基于卷积神经网络的端到端多维语音识别模型流程图

经过文献查阅和实验验证，参考了诸如 VGGNet 等经典 CNN 网络的结构，本章模型中的 CNN 结构设置如图 4.6 所示。共设置 5 个卷积层、5 个池化层和 1 个全连接层，其中每个卷积层的 padding 参数都为 same 模式，保证卷积之后的特征尺寸保持不变；每个卷积层的输出值都经过一个激活函数，这里采用最常用的 ReLU 函数，接着再输出到下一层；另外，在每个卷积层中采取了 batch-normalization 来加速训练，增强泛化能力，防止过拟合。具体每层的参数设置如表 4.1 所示。

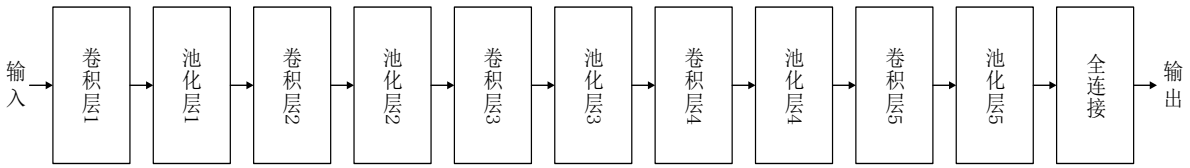

图 4.6 本文 CNN 结构图

表 4.1 本文 CNN 具体参数设置

名称	卷积/池化尺寸	步长
卷积层 1	16 个 3×3	1
池化层 1	1×2	2
卷积层 2	32 个 3×3	1
池化层 2	1×2	2
卷积层 3	32 个 3×3	1
池化层 3	1×2	2
卷积层 4	64 个 3×3	1
池化层 4	1×2	2
卷积层 5	64 个 3×3	1
池化层 5	1×2	2

最后经过一个全连接层使得 CNN 最终的输出中每个时间步对应的特征为 32 维。随后将得到的特征向量按顺序输入到后续的 RNN 多任务网络中进行分类。与第三章模型类似，RNN 多任务网络设置三层 RNN 作为共享层，采用 GRU 作为网络基本单元，每层 512 个节点；共享层之后设置三个单独的全连接属性依赖层，各 512 个节点；最后属性依赖层的输出通过三个 softmax 层输出三项识别任务的分类结果。模型训练和测试的方式都和第三章相同，训练阶段通过计算损失、优化器优化损失函数来更新网络权重，测试阶段测试模型性能。模型损失函数的设定和第三章一样采用交叉熵函数，各识别任务损失函数的权重选择也同样通过验证集来进行调试和确定。

4.4 基于卷积神经网络和特征融合的多维语音识别模型

4.4.1 特征融合

在 4.3 节中提出的基于 CNN 的端到端多维语音识别模型通过 CNN 对语谱图进行特征提取，得到用于多维识别的语音特征参数，其过程并未有人工干预，由 CNN 自行训练完成，相对于传统的 MFCC 等人工特征，基于 CNN 提取的特征从不同的角度对语音信号中的信息进行了表征，同时也尽可能多的保留了语音信号中的多维度信息。但是，从另一个角度来看，在基于 CNN 的特征提取算法模型的构建中也存在着一些问题。首先，诸如 MFCC 等人工特征中包含了过往研究的先验知识，而基于 CNN 提取的特征由于是由网络来自行学习，故没有这个优势，因为 CNN 训练时并不了解其输出结果需要作为多任务分类器的输入，获得最佳多任务分类的目的，仅仅是从输入表征上考虑问题；其次，CNN 网络的构成没有一成不变的定式，需要根据识别任务的类型、语音库的情况等多方面来设计，其结构的设计以及训练和测试中的各项超参数的设定没有办法确保最优，可能存在不完善的问题；最后，在网络模型进行实际应用时，可能会出现一些无法预料的问题，例如在语谱图进行尺寸统一的时候，可能会丢失部分信息，对模型最终的识别结果产生影响。

以往的研究表明，单一类别的特征很难全面的表达样本信息，因此过往的很多研究选择特征融合的方式，从多个角度更加全面的对语音信号中的信息进行描述[68-71]。因此本章最终识别模型输入参数的选择考虑采用特征融合的方法，将基于 CNN 提取的特征和 MFCC 特征进行融合拼接，形成融合特征。一方面，MFCC 特征在提取过程中的滤波等各种信号处理方式会使得其丢失很多有用的特征，而基于 CNN 提取的特征中包含大量的原始语音信息，可以对 MFCC 特征进行补充；另一方面，MFCC 特征是经过长时间研究检验的有效的人工特征，

包含了有用的先验知识，可以为 CNN 特征提供性能保证和补充，在一定程度上解决上文描述的问题。采用特征融合的方式可以结合两种特征的优势，达到二者互补的目的，提高特征的有效性和多维识别的准确率，也更加适合多维语音识别。

4.4.2 算法模型设计

本节研究提出了一种基于卷积神经网络和特征融合的多维语音识别算法，其整体系统流程图如图 4.7 所示。将语音信号的 WAV 文件输入到模型中，经过与上文相同的语音活动检测、重采样、分帧等操作后，一方面转化为语谱图输入到 CNN 中进行特征提取，其中 CNN 网络结构的各项参数和设定均与 4.3.2 节中相同，得到 32 维特征；另一方面对其提取 MFCC 特征，提取方法与第三章相同，最终输出 40 维组合 MFCC 特征。接下来将 MFCC 特征和基于 CNN 的特征进行归一化，随后进行融合拼接，得到 72 维融合特征，拼接方式如式（4.2）所示。

$$V^n = [aV_1^n, bV_2^n] \tag{4.2}$$

其中，V_1^n 和 V_2^n 分别表示第 n 帧的基于 CNN 提取的特征和组合 MFCC 特征，在分别乘以权重 a 和 b 之后进行向量拼接。a 和 b 的和为 1，其具体取值在仿真实验中通过验证集进行调试和确定。

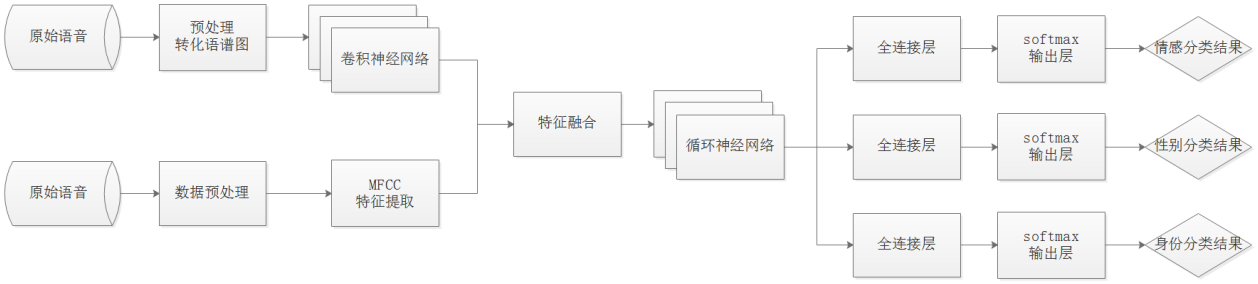

图 4.7 基于卷积神经网络和特征融合的多维语音识别流程图

最终，将拼接完成后的融合特征输入到后续的 RNN 多任务网络中进行分类，输出三项识别任务的结果。其中 RNN 多任务网络的结构和参数设置以及训练和测试的流程与 4.3 节中的设定完全一致，这里不再赘述。

4.5 实验结果分析

4.5.1 实验语音数据库

由于本章所构建的多维识别算法模型的网络结构层数较多，故其仿真实验训练中对所用

的实验语音数据库有一定的样本数量要求，若样本数据量太少则可能不足以充分训练网络模型。因此本文采取了与第三章的研究中相同的CASIA和KSU两个语音数据库中全部的数据，一方面这两个数据库的样本数据量较大；另一方面也方便将本章的算法模型和第三章的算法模型进行性能上的比较。其中 CASIA 中包含 2 男 2 女共 4 个说话人，六种情感（anger，fear，happy，neutral，sad，surprise），共 7200 条语音数据；KSU 包含 7 男 7 女共 14 个说话人，五种情感（anger，happy，neutral，sad，surprise），共 1680 条语音数据。本章在仿真实验中同样将两个数据库中的语音样本数据分别以 6:2:2 的比例分为训练集、验证集和测试集，分别用于模型的训练、各项超参数的确定、识别性能的评估。

4.5.2 实验环境参数设置

本章的实验在 Python 平台上进行，使用 librosa 语音信号处理工具包对语音进行预处理、MFCC 特征提取、语谱图转化；使用基于 Python3.6 的 GPU 版本 TensorFlow 框架进行神经网络模型的搭建、训练和测试。在训练阶段，模型采用 mini-batch ADAM 优化算法对总损失函数进行优化，批次大小设置为 30；每 50 个批次测试一次每项任务的训练集识别准确率和测试集识别准确率；网络模型中 CNN 部分的激活函数采用 ReLU 函数，输出 softmax 层采用 softmax 函数，其余部分均采用 tanh 函数；基础学习率设定为 0.001，衰减系数为 0.99，每 50 个批次衰减一次；另外，网络设定了 0.5 的 dropout，并采用了 batch-normalization，目的是加速训练，增强泛化能力，防止梯度消失和过拟合。

本章同样通过验证集来确定模型总损失函数中各项任务损失函数权重系数 α、β、γ 的取值，以及特征融合系数 a 和 b 的取值。以 0.05 为间隔对 a 和 b 进行搜索，固定 a 和 b 后使用同第三章的方法在验证集上测试性能得到当前 a 和 b 取值下的最高识别率和 α、β、γ 的取值，再调整 a 和 b 的取值进行同样的操作，最终得到使模型在验证集上取得最高识别率时对应的 a、b 以及 α、β、γ 的值，分别为 0.45、0.55、0.5、0.15、0.35（CASIA）以及 0.4、0.6、0.25、0.1、0.65（KSU）。此处 α 和 γ 的取值在两个数据库中存在差异，主要是由于损失函数权重系数的取值与数据库对应的识别任务关联较大。在 CASIA 中，α 对应的情感识别任务为 6 选 1，γ 对应的身份识别任务为 4 选 1，而在 KSU 中情感识别任务为 5 选 1，身份识别任务为 14 选 1，在识别任务上二者有较大差别。在实际应用中也需要根据识别任务的情况进行验证集实验来确定参数取值。

4.5.3 对比实验设置

为检验本章提出的基于卷积神经网络和特征融合的多维语音识别模型的性能表现，设置以下仿真实验：

（1）首先对本章 4.3 节和 4.4 节中提出的仿真模型进行实验，分别是基于 CNN 的端到端多维语音识别模型以及基于 CNN 和特征融合的多维语音识别模型，分别记做 MTL-CNN 和 MTL-Fusion，记录二者的各项任务识别率。

（2）接着为了验证多任务模型相对于单维语音识别模型在性能上的提升，分别针对（1）中的两个模型设置单维语音识别对比实验，分别记做 STL-CNN 和 STL-Fusion，在保持模型结构和超参数相同的条件下，分别对说话人情感、身份、性别三项任务进行单维识别，并记录各项识别任务的识别率。

（3）随后将第四章的多维语音识别模型与第三章中采用 MFCC 作为语音特征参数的基于多任务学习和循环神经网络的多维语音识别模型 MTL-RNN 进行性能比较，这里不需要进行重复实验，直接使用 3.5.4 中的实验结果进行对比即可。

（4）最后为测试本章提出的 MTL-Fusion 模型在噪声环境下的性能表现，设计几组仿真对比实验，将 KSU 语音库中用于测试的部分数据人为加入指定信噪比的高斯白噪声，之后将混合测试信号输入到模型中，最后输出噪声下的各项任务识别率。其中信噪比的值设定为 0dB 到 35dB，共 8 组实验，记录并对比实验结果，显示本章模型的抗噪性能。

4.5.4 结果分析

（1）基于卷积神经网络的端到端多维识别模型与单维识别模型性能比较

首先将本章 4.3 节中提出的基于卷积神经网络的端到端多维识别模型和相同 CNN 结构、相同条件下的单维识别模型进行性能比较。表 4.2 显示了单维和多维模型在 CASIA 和 KSU 两个语音数据库上三项识别任务的识别准确率。

表 4.2 基于 CNN 的单维模型和多维模型在 CASIA 和 KSU 上的各项任务识别率

语音库	模型	性别	身份	情感	平均
CASIA	MTL-CNN	99.86	99.31	88.74	95.97
	STL-CNN	99.20	97.75	81.39	92.78
KSU	MTL-CNN	99.14	93.69	87.62	93.48
	STL-CNN	98.56	87.36	78.48	88.13

由表 4.2 可以看出，本章 4.3 节中提出的基于卷积神经网络的端到端多维语音识别模型在两个语音库上的各项任务识别率相对于单维识别有不同程度的性能提升，平均识别率分别提升了 3.19%（CASIA）和 5.35%（KSU）。由于 CASIA 语音库中说话人数量较少，故多维识别在 CASIA 语音库上的性能提升主要体现在情感识别上；而在 KSU 语音库中情感和身份识别的识别率均有明显的提升。实验结果再次证明了多维语音识别的有效性以及基于 CNN 提取的语音特征参数用来进行多维语音识别的可行性。

（2）基于融合特征的多维识别模型与单维识别模型性能比较

接着将本章 4.4 节中提出的基于卷积神经网络和特征融合的多维语音识别模型和相同 CNN 结构、相同条件、相同 MFCC 特征的单维识别模型进行性能比较。表 4.3 显示了单维和多维模型在 CASIA 和 KSU 两个语音数据库上各项识别任务的识别准确率。

表 4.3 基于融合特征的单维模型和多维模型在 CASIA 和 KSU 上的各项任务识别率

语音库	模型	性别	身份	情感	平均
CASIA	MTL-Fusion	99.99	99.56	92.95	97.50
	STL-Fusion	99.38	98.41	83.95	93.91
KSU	MTL-Fusion	99.96	96.97	92.06	96.33
	STL-Fusion	99.04	89.88	82.05	90.32

从表中数据可以看出，本章 4.4 节中提出的基于卷积神经网络和特征融合的多维语音识别模型在两个语音库上的各项任务识别率相对于单维识别同样均有提升，平均识别率分别提升了 3.59%（CASIA）和 6.01%（KSU），大于（1）中 CNN 端到端模型的多维平均识别率提升，表明融合特征更加符合多维语音识别的特征参数要求。同样的，融合特征多维识别在 CASIA 语音库上的性能提升主要体现在情感识别上，而在 KSU 语音库上各项任务均有明显性能提升。实验证明了基于 CNN 提取的特征和 MFCC 的融合特征用于多维语音识别的有效性。

（3）本章基于 CNN 的模型与第三章 MFCC 模型的性能比较

为了显示本章提出的基于卷积神经网络和特征融合的多维语音识别模型 MTL-Fusion 的性能优越性，将其与本文第三章中的基于多任务学习和循环神经网络的多维语音识别模型 MTL-RNN 以及本章 4.3 节中的基于 CNN 的端到端多维语音识别模型 MTL-CNN 进行性能比较，实验结果如表 4.4 所示，表中展示了三种模型在两个语音库上的各项任务识别率、平均识别率以及多维识别相对于单维识别的识别率提升。

表 4.4 本文三种模型的各项任务识别率比较

语音库	模型	性别	身份	情感	平均	提升
CASIA	MTL-RNN	99.99	99.09	90.87	96.65	3.01
	MTL-CNN	99.86	99.31	88.74	95.97	3.19
	MTL-Fusion	99.99	99.56	92.95	97.50	3.59
KSU	MTL-RNN	99.70	95.83	90.48	95.34	5.09
	MTL-CNN	99.14	93.69	87.62	93.48	5.35
	MTL-Fusion	99.96	96.97	92.06	96.33	6.01

由表 4.4 中数据可以看出，MTL-CNN 的模型性能在说话人性别识别和身份识别上与 MTL-RNN 相差不大，在说话人情感识别上略差于 MTL-RNN，说明单一的通过 CNN 提取的特征并没有取得非常好的效果；MTL-Fusion 在各项任务的识别率上均高于 MTL-CNN 和 MTL-RNN，这也印证了特征融合策略可以使基于 CNN 提取的特征和 MFCC 特征互补，可有效提高模型性能；在多维识别相对于单维识别的性能提升上，采用融合特征的 MTL-Fusion 要高于采用 MFCC 特征的 MTL-RNN 和采用基于 CNN 提取的特征的 MTL-CNN，说明融合特征将从不同角度描述语音信号信息的两种特征进行融合的方式能够更好的利用到语音信号中所包含的丰富的多维度信息，对相关任务间的相关信息有更好的表征，更加符合多维语音识别的语音特征参数要求。与 MTL-RNN 模型对比来看，MTL-Fusion 模型在两种语音库上的平均识别率分别提升了 0.85%和 0.99%，在 CASIA 库中的识别率提升主要集中在情感识别上，在 KSU 库中身份和情感上的识别率均有提升，证明 MTL-Fusion 模型具有更好的性能。

（4）基于融合特征的模型在噪声环境下的性能表现

如第三章中的仿真实验一样，为了测试本章提出的基于融合特征的模型在噪声环境下的性能表现，以同样的方式保持 KSU 数据库的训练集和验证集不变，向测试集中人为添加指定信噪比的高斯白噪声，测试模型的各项任务性能，并与 MTL-RNN 模型在噪声下的性能进行对比。实验结果如表 4.5 所示。

表 4.5 MTL-Fusion 和 MTL-RNN 在不同信噪比的噪声环境下的性能

信噪比	模型	性别	身份	情感	平均
0dB	MTL-RNN	77.34	22.31	44.74	48.13
	MTL-Fusion	79.25	25.63	44.42	49.77
5dB	MTL-RNN	88.26	29.65	60.70	59.54
	MTL-Fusion	89.34	32.34	62.36	61.35
10dB	MTL-RNN	90.14	42.23	64.68	65.68
	MTL-Fusion	90.78	44.71	67.42	67.64
15dB	MTL-RNN	96.12	52.07	67.33	71.84
	MTL-Fusion	97.03	55.89	69.97	74.30
20dB	MTL-RNN	98.34	61.61	73.34	77.76
	MTL-Fusion	98.32	63.70	75.48	79.17
25dB	MTL-RNN	98.65	72.54	76.37	82.52
	MTL-Fusion	98.70	79.24	79.62	85.85
30dB	MTL-RNN	98.81	90.44	82.12	90.46
	MTL-Fusion	98.89	92.35	86.65	92.63
35dB	MTL-RNN	99.41	95.73	89.01	94.72
	MTL-Fusion	99.47	96.88	91.73	96.03
无噪声	MTL-RNN	99.70	95.83	90.48	95.34
	MTL-Fusion	99.96	96.97	92.06	96.33

由表中数据可以看出，两个模型的平均识别率变化趋势相同，均随着信噪比的增加不断提升，在信噪比高于 30dB 时模型平均识别率接近无噪声状态下的识别率值。MTL-Fusion 模型在每个信噪比下的平均识别率均高于 MTL-RNN 模型，且相对更接近于无噪声模型性能。实验证明 MTL-Fusion 模型具有更好的抗噪性能，因此也表明特征融合有助于提高系统的鲁棒性，当然，如果需要继续提升系统抗噪性能，需要在前端输入信号部分采用消噪技术降低噪声影响，也是本文后续研究方向。

（5）系统运行时间比较

对本文第三章的 MTL-RNN 模型和第四章的 MTL-Fusion 模型以及作性能比较的文献[3]中的模型进行系统运行时间的比较，记录训练后的模型对整个测试集共 1440 条数据进行识别所用的时间（包括语音特征提取的时间），结果如表 4.6 所示。

表 4.6 系统运行时间比较

模型	识别时间
MTL-RNN	82.8s
MTL-Fusion	159.5s
文献[3]	121.2s

由表中数据可以看出，MTL-Fusion 模型在取得性能提升的同时，也由于模型复杂度的增加使得系统进行识别所需要的时间大于其他两个模型，识别时间增加的幅度尚在可接受范围内，反映了模型算法的可行性。

4.6 本章小结

本章提出了一种采用卷积神经网络和特征融合方法的多维语音识别算法，采用卷积神经网络进行特征提取，并将基于 CNN 的特征和 MFCC 特征进行特征融合，之后输入到带有属性依赖层的多任务循环神经网络分类器中，最后同时输出说话人身份、情感、性别三项识别任务的分类结果，在特征提取上充分利用了语音信号中丰富的多维度信息，有效提升了多维语音识别模型的分类性能。本章首先介绍了卷积神经网络的概念和原理，接着构建了基于 CNN 的端到端多维语音识别模型，随后将 CNN 提取的特征和 MFCC 进行特征融合，构建了基于融合特征的多维语音识别模型。最后，对本章提出的 MTL-Fusion 模型以及各项对比实验进行仿真实验，实验结果表明本章提出的基于卷积神经网络和特征融合的多维语音识别模型在两种语音库上的平均识别率分别为 97.5%和 96.33%，比单维识别高出 3.59%和 6.01%，比第三章中使用 MFCC 作为语音特征参数的基于多任务学习和循环神经网络的多维语音识别模型高出 0.85%和 0.99%，在三项识别任务的识别率上均有提升，且具有更好的抗噪性能。

第五章 总结与展望

5.1 论文总结

随着信息技术的高速发展，5G 时代正在到来，在人工智能的浪潮下，实现简单、快捷、流畅的人机交互成为人类追求的目标，而语音识别技术正是人机交互领域中重要的组成部分。近年来，不少优秀的研究者在语音识别领域做了许多工作，其成果也颇为丰硕，但就目前的语音识别研究而言，主要以单维识别为主。为了使语音识别技术能够更加拟人化、智能化，即计算机能够像人一样对语音信号进行识别和分析，本团队提出了对语音信号中的多维信息同时进行识别的课题，充分利用语音信号中所包含的不同语音信息之间的相关性，对多项语音识别任务进行同时分类，最终目的是实现智能化人机语音交互。本文在本团队前期研究的基础上继续深入研究多维语音识别技术，侧重在继续提高多任务系统各任务识别性能，主要研究的是说话人性别、情感、身份三类语音信息的同时识别，采用循环神经网络、卷积神经网络结合多任务学习理论实现语音信号的多维识别。本文的主要工作和创新点如下：

（1）为了充分利用语音信号中丰富的多维信息和不同识别任务间的相关信息，实现多维语音信息的同时识别，本文将多任务学习机制与循环神经网络结构相结合，构建了基于 MTL-RNN 的多维语音识别模型，实现了说话人性别、情感、身份的同时识别。语音信号经过预处理后进行 MFCC 特征提取并输入到模型中，利用带有属性依赖层的多任务神经网络结构，通过 RNN 共享层共享神经网络参数以学习各识别任务间的共有特征；通过全连接属性依赖层分别学习各识别任务的独有特征；通过调整模型总损失函数中各识别任务损失函数的权重占比来针对不同类型的语音数据库进行模型性能优化；最终同时输出三种识别任务的识别结果。最后对 MTL-RNN 模型进行实验验证，通过多项对比实验测试模型性能、结构有效性以及抗噪性能。实验结果表明，本文提出的基于多任务学习和循环神经网络的多维语音识别模型有较好的性能，在 CASIA 和 KSU 语音库上的平均识别率为 96.65%和 95.34%，分别比单维识别高出 3.01%和 5.09%，三项识别任务均有一定的识别率提升，对于语种因素和说话人的个性因素有较好的鲁棒性，且具备一定的抗噪性能。

（2）由于第三章中的 MTL-RNN 模型采用的语音特征是常规的 MFCC 特征，可能已经失去部分语音原始信息，并不一定最适合多维语音识别，因此为了对多维识别语音特征提取部分进行改进，本文将卷积神经网络结构与特征融合方法相结合，进行语音特征参数的提取，将语音信号原始语谱图经过 CNN 提取的特征和人工提取的 MFCC 特征进行融合，使用融合

特征输入到多任务循环神经网络分类器中进行说话人身份、性别、情感三项任务的识别，构建了基于 CNN 和特征融合的多维语音识别模型 MTL-Fusion，在特征提取上充分利用了语音信号中的多维度信息，使两种特征进行互补，有效提升了多维语音识别模型的分类性能。最后对 MTL-Fusion 模型进行实验验证，实验结果证明了融合特征在多维语音识别上的有效性，本文提出的基于卷积神经网络和特征融合的多维语音识别模型，在 CASIA 和 KSU 语音库上的平均识别率为 97.5%和 96.33%，分别比单维识别高出 3.59%和 6.01%，比非特征融合的 MTL-RNN 模型高出 0.85%和 0.99%，三项识别任务均有识别率提升，且具有更好的抗噪性能。

5.2 工作展望

本文主要研究了多维语音识别的实现，使用多任务学习理论、卷积神经网络和循环神经网络，从分类模型构建和特征提取两个方面对多维语音识别进行研究，提出了基于多任务学习和循环神经网络的多维语音识别模型，以及基于卷积神经网络和特征融合的多维语音识别模型，证明了多维语音识别的可实现性和有效性。由于多维语音识别是一个新颖的课题，目前的研究较少，本团队也只是处于起步阶段，加上本人的研究水平、代码能力和时间均有限，语音识别训练需要的较大的数据量等因素，故本文仍然有许多可以改进和拓展之处，多维语音识别研究距离真正的拟人化人机语音交互，还有很多方面需要进行更加深入的研究。

（1）为了更好的从单维识别研究中学习和吸取经验，本文的研究中选取了技术较为成熟的说话人身份、性别、情感识别作为多维识别的三项识别任务，而语音信号包含的丰富信息远不止这三类信息，三种任务同时识别与真正的拟人化语音识别之间还有很长的距离，未来的多维语音识别研究需要继续考虑更多的其他种类的识别任务，例如语义识别，说话人年龄识别，语种和方言识别，甚至环境背景声音识别等识别任务，实现更多维的识别。

（2）本文的两个模型所采用的特征参数分别是 MFCC 以及基于 CNN 提取的特征与 MFCC 的融合特征，虽然融合特征的效果相比 MFCC 已经有所提升，但仍然有较大的改进空间。例如，本文的融合是比较简单的，未来的研究可以考虑更多其他类型特征的融合，例如韵律特征、i-vector 特征等，选择更加复杂的融合方式，考虑特征冗余和降维。当然，也可以另辟蹊径寻找其他适合多维语音识别的特征。

（3）本文选择了模型结构较为复杂的神经网络来进行多维语音识别，在模型训练的效率方面有些欠缺。未来的研究可以从提高模型分类效率、降低复杂度方面入手，可参考谷歌、科大讯飞等前沿科技公司的建模方法，优化网络结构，提高训练效率，当然这也需要很好的代码能力支撑。

（4）多维语音识别可以引申到多模态的语音识别，不局限于语音中的多任务，引入其他类别的相关任务，例如语音和图像结合，通过语音和唇形结合判断内容、语音和面部表情结合判断身份或者情感等，或者是语音识别和自然语言处理结合，如话语情感的判断可以辅助语义的判断等。这与多维识别的概念相契合，也对拟人化人机交互的实现有所帮助。

参考文献

[1] 赵力. 语音信号处理[M]. 机械工业出版社, 2003.
[2] 李姗. 多维语音信息识别技术研究[D]. 南京邮电大学学位论文, 2017.
[3] 陈海霞. 基于神经网络的多维说话人信息识别研究[D]. 南京邮电大学学位论文, 2018.
[4] 赵力, 黄程韦. 实用语音情感识别中的若干关键技术[J]. 数据采集与处理, 2014, 29(2):157-170.
[5] 陈海霞, 徐珑婷, 杨震. 渐进式神经网络多维说话人信息识别技术[J]. 南京邮电大学学报：自然科学版, 2019, 39(1):45-51.
[6] Chen H, Xu L, Yang Z. Multi-dimensional speaker information recognition with multi-task neural network[C]. IEEE International Conference on Computer and Communications (ICCC 2018).
[7] Li S, Xu L, Yang Z. Multidimensional speaker information recognition based on proposed baseline system[C]. 2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). IEEE, 2017.
[8] 郑方, 王仁宇, 李蓝天. 生物特征识别技术综述[J]. 信息安全研究, 2016, 2(1): 12-26.
[9] Chen W, Hong Q, Li X. GMM-UBM for text-dependent speaker recognition[C]. International Conference on Audio, Language and Image Processing. 2012:432-435.
[10] Kestra L G. Voiceprint Identification[J]. Nature, 1962, 34(5): 1253-1257.
[11] Quatieri T F(著). 赵胜辉, 刘家康等(译). 离散时间语音信号处理——原理与应用[M]. 电子工业出版社, 2004.
[12] O'Shaughnessy D. Speaker recognition[J]. IEEE Acoustic, Speech and Signal Processing Magazine. 1986, (4):4-17.
[13] 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): 1304-1312.
[14] Keogh E, Ratanamahatana C A. Exact indexing of dynamic time warping[J]. Knowledge and information systems, 2005, 7(3): 358-386.
[15] Soong F K, Rosenberg A E, Juang B H, et al. Report: A vector quantization approach to speaker recognition[J]. AT&T technical journal, 1987, 66(2): 14-26.
[16] Mohanty S, Swain B K. Speaker Identification using SVM during Oriya Speech Recognition[J]. International Journal of Image Graphics & Signal Processing, 2015, 7(10):28-36.
[17] Farrell K R, Mammone R J, Assaleh K T. Speaker recognition using neural networks and conventional classifiers[J]. IEEE Transactions on speech and audio processing, 1994, 2(1): 194-205.
[18] 薛少飞. DNN-HMM 语音识别声学模型的说话人自适应[D]. 中国科学技术大学, 2015.
[19] Childers D G, Wu K, Bae K S, et al. Automatic recognition of gender by voice[C]. International Conference on Acoustics. IEEE, 1988.
[20] 陈国杰. 状态变量带通滤波器及其在男女声识别中的应用[J]. 电声技术, 2003(06):46-48.
[21] 吴朝晖. 说话人识别模型与方法[M]. 清华大学出版社, 2009.
[22] Chen O T C, Gu J J, Lu P T, et al. Emotion-inspired age and gender recognition systems[C]. Circuits and Systems (MWSCAS), 2012 IEEE 55th International Midwest Symposium on. IEEE, 2012: 662-665.
[23] 贺文锋. 说话人性别识别与年龄估计的研究[D]. 华南理工大学, 2014.
[24] 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.
[25] Lin Y L, Wei G. Speech emotion recognition based on HMM and SVM[C]. International Conference on Machine Learning and Cybernetics. IEEE, 2005:4898-4901 Vol. 8.
[26] Hu H, Xu M X, Wu W. GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2007:IV-413 - IV-416.

[27] Schuller B, Rigoll G, Lang M. Hidden Markov model-based speech emotion recognition[C]. International Conference on Multimedia and Expo, 2003. ICME '03. Proceedings. IEEE, 2003:I-401-4 vol.1.
[28] Parthasarathy S, Busso C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning[C]. INTERSPEECH. 2017:1103-1107.
[29] Harár P, Burget R, Dutta M K. Speech emotion recognition with deep learning[C]. International Conference on Signal Processing and Integrated Networks. IEEE, 2017:137-140.
[30] Huang Z, Dong M, Mao Q, et al. Speech Emotion Recognition Using CNN[C]. Acm International Conference on Multimedia. ACM, 2014:801-804.
[31] Yin X, Liu X. Multi-task convolutional neural network for pose-invariant face recognition[J]. IEEE Transactions on Image Processing, 2018, 27(2): 964-975.
[32] Xia R, Liu Y. A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space[M]. IEEE Computer Society Press, 2017.
[33] 吕亮. 基于深度学习的说话人识别方法的研究[D]. 东南大学学位论文, 2016.
[34] 马运杰. 基于稀疏表示的鲁棒性说话人识别技术研究[D]. 南京邮电大学学位论文, 2015.
[35] 周伟栋. 语音压缩感知系统中的消噪技术研究[D]. 南京邮电大学学位论文, 2016.
[36] 尹栋, 蒋涉权, 刘宝光等. 语音增强算法综述及性能分析[J]. 电声技术, 2015, 39(5):58-61.
[37] 谷鹏. 压缩采样系统中的语音消噪技术研究[D]. 南京邮电大学学位论文, 2014.
[38] Estevez P A, Becerra-Yoma N, Boric N, et al. Genetic programming-based voice activity detection[J]. Electronics Letters, 2005, 41(20):1141-1143.
[39] 蒋晔. 基于短语音和信道变化的说话人识别研究[D]. 南京理工大学学位论文, 2012.
[40] 杨迪, 戚银城, 刘明军, 等. 说话人识别综述[J]. 电子科技, 2012, 25(6): 162-165.
[41] 金碧程. 基于深度学习的语音情感识别研究[D]. 北京邮电大学学位论文, 2018.
[42] 孙亚新. 语音情感识别中的特征提取与识别算法研究[D]. 华南理工大学学位论文, 2015.
[43] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1):1929-1958.
[44] Soong F K, Rosenberg A E. On the use of instantaneous and transitional spectral information in speaker recognition[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1988, 36(6):871-879.
[45] Davis S V, Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences[J]. IEEE Transactions on Acoustics Speech and Signal Processing, 1980, 28(4):57-366.
[46] Kinnunen T, Li H. An overview of text-independent speaker recognition: From features to supervectors[J]. Speech Communication, 2010, 52(1):12-40.
[47] Xu L, Yang Z, Sun L. Simplification of I-Vector Extraction for Speaker Identification[J]. Chinese Journal of Electronics, 2016, 25(6):1121-1126.
[48] 马平, 黄浩, 程露红, 等. 基于i-vector说话人识别算法中训练时长研究[J]. 现代电子技术, 2016, 39(14):1-3.
[49] 李姗, 徐珑婷. 基于语谱图提取瓶颈特征的情感识别算法研究[J]. 计算机技术与发展, 2017(5).
[50] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation. 1997, 9(8):1735-1780.
[51] Graves A, Mohamed A R, Hinton G. Speech Recognition with Deep Recurrent Neural Networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013.
[52] Kim K, Lee J, Ha H, etc. Speech emotion recognition based on multi-task learning using a convolutional neural network[C]. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).
[53] Tzirakis P, Zhang J, Schuller B. End-to-End Speech Emotion Recognition Using Deep Neural Networks[C]. International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2015.
[54] 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 (ICASSP). IEEE, 2015.
[55] Caruana R A. Multitask Learning: A Knowledge-Based Source of Inductive Bias[J]. Machine Learning Proceedings, 1993, 10(1):41-48.
[56] Chung J, Gulcehre C, Cho K H, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[J]. Eprint Arxiv, 2014.
[57] Schmidt M, Gish H. Speaker identification via support vector classifiers[C]. Acoustics, Speech, and Signal Processing, 1996. on Conference Proceedings. 1996 IEEE International Conference. IEEE Computer Society, 1996:105-108.
[58] Caruana R. Multitask Learning[J]. Machine Learning, 1997, 28(1):41-75.
[59] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2):157-166.
[60] Sciarrone A, Delfino A, Marchese M, et al. Gender-driven emotion recognition through speech signals for ambient intelligence applications[J]. IEEE Transactions on Emerging Topics in Computing, 2014, 1(2):244-257.
[61] Mefiah A, Alotaibi Y A, Selouani S A. Arabic speaker emotion classification using rhythm metrics and neural networks[C]. Signal Processing Conference (EUSIPCO), 2015 23rd European. IEEE, 2015: 1426-1430.
[62] Goodfellow I, Bengio Y, Courville A. Deep Learning[M]. 人民邮电出版社, 2017.
[63] Lecun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989, 1(4):541-551.
[64] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324.
[65] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).
[66] Simonyan K , Zisserman A . Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. Computer Science, 2014.
[67] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. 2015.
[68] 卢官明, 袁亮, 杨文娟等. 基于长短期记忆和卷积神经网络的语音情感识别[J]. 南京邮电大学学报(自然科学版), 2018, 38(05):67-73.
[69] 李鹏程. 基于深度学习的语音情感识别研究[D]. 中国科学技术大学学位论文, 2019.
[70] Jin Y , Song P , Zheng W , et al. A feature selection and feature fusion combination method for speaker-independent speech emotion recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2014.
[71] 张雄. 基于卷积神经网络特征优化的语音情感识别研究[D]. 华中师范大学学位论文, 2018.
[72] Baxter J. A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling[J]. Machine Learning, 1997, 28(1):7-39.
[73] Li R, Zhao M, Li Z, etc. Anti-Spoofing Speaker Verification System with Multi-Feature Integration and Multi-Task Learning[C]. INTERSPEECH. 2019:1048-1052.
[74] Sarria-paja M , Senoussaoui M , O'Shaughnessy D , et al. Feature Mapping, Score-, and Feature-Level Fusion for Improved Normal and Whispered Speech Speaker Verification[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.
[75] You L, Guo W, Dai L, etc. Multi-Task Learning with High-Order Statistics for X-vector based Text-Independent Speaker Verification[C]. INTERSPEECH. 2019:1158-1162.
[76] Guo L, Wang L, Dang J, et al. A Feature Fusion Method Based on Extreme Learning Machine For Speech Emotion Recognition[C]. IEEE International Conference on Acoustics, Speech and Signal Processing

(ICASSP). IEEE, 2018.

[77] Kurata G, Audhkhasi K. Multi-task CTC Training with Auxiliary Feature Reconstruction for End-to-end Speech Recognition[C]. INTERSPEECH. 2019:1636-1640.

[78] Mendes C. Abad A, Neto P, etc. Recognition of Latin American Spanish using Multi-task Learning[C]. INTERSPEECH. 2019:2135-2139.

[79] 张仕良. 基于深度神经网络的语音识别模型研究[D]. 中国科学技术大学学位论文, 2017.

[80] Tavarez D, Sarasola X, Alonso A, et al. Exploring Fusion Methods and Feature Space for the Classification of Paralinguistic Information[C]. INTERSPEECH 2017.

[81] Xu S, Fosler-Lussier E. Application of Progressive Neural Networks for Multi-Stream Wfst Combination in One-Pass Decoding[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5914-5918.

[82] Xu L, Lee K A, Li H, et al. Generalizing I-Vector Estimation for Rapid Speaker Recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2018:1-1.

[83] Zhang Y, Liu Y, Weninger F, et al. Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017:4990-4994.

[84] Jati A, Peri R, Pal M, etc. Multi-task Discriminative Training of Hybrid DNN-TVM Model for Speaker Verification with Noisy and Far-Field Speech[C]. INTERSPEECH. 2019:2463-2467.

[85] Gowda D, Garg A, Kim K, etc. Multi-task multi-resolution char-to-BPE cross-attention decoder for end-to-end speech recognition[C]. INTERSPEECH. 2019:2783-2787.

[86] Chang J, Scherer S. Learning representations of emotional speech with deep convolutional generative adversarial networks[C]. IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2017:2746-2750.

[87] Liu G, He W, Jin B, et al. Feature Fusion of Speech Emotion Recognition Based on Deep Learning[C]. 2018 6th IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC). IEEE, 2018.

[88] Pan H, Li X, Huang Z. A Mandarin Prosodic Boundary Prediction Model Based on Multi-Task Learning[C]. INTERSPEECH. 2019:4485-4488.

[89] Sebastian J, Pierucci P. Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts[C]. INTERSPEECH. 2019:51-55.

[90] Tralie C, McFee B. Enhanced Hierarchical Music Structure Annotations Via Feature Level Similarity Fusion[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 201-205.

[91] Maeda K, Takahashi Sho, Ogawa T. Multi-feature Fusion Based on Supervised Multi-view Multi-label Canonical Correlation Projection[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 3936-3940.

[92] Manjunath K, Rao K, Jayagopi D, etc. Indian languages ASR: A multilingual phone recognition framework with IPA based common phone-set, predicted articulatory features and feature fusion[C]. INTERSPEECH. 2018:1016-1020.

[93] Zhou Z H, Zhang M L, Huang S J, et al. MIML: A Framework for Learning with Ambiguous Objects[J]. Corr Abs, 2008, abs/0808.3231(1):2012.

[94] Chen D, Mak K W. Multitask learning of deep neural networks for low-resource speech recognition[J]. IEEE ACM Transactions on Audio Speech & Language Processing, 2015, 23(7):1172-1183.

[95] Toyoda A, Ogawa T, Haseyama M. Semi-supervised Multiple Feature Fusion for Video Preference Estimation[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 891-895.

[96] Liu N, Fang Y, Li L, etc. Multiple Feature Fusion for Automatic Emotion Recognition Using EEG Signals[C]. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 896-900.

[97] Amodei D, Anubhai R, Battenberg E, et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin[J]. Computer Science, 2015.

附录 1 攻读硕士学位期间撰写的论文

[1] 冯天艺, 杨震. 采用多任务学习和循环神经网络的语音情感识别算法[J]. 信号处理, 2019，35(7).

附录 2 攻读硕士学位期间参加的科研项目

（1）国家自然科学基金“鲁棒性压缩感知关键技术的研究”，项目编号：61271335

（2）国家“863”高技术研究发展计划重点项目“多语言语音识别关键技术研究与应用产品开发”，项目编号：2006AA010102

致谢

时光匆匆，转眼间，三年的研究生生活已接近尾声。回首看，师门第一次开会的场景仿佛还在昨天。三年里，导师、同门、同学们的诸多帮助让我能够顺利的在研究生道路上稳步前进。在论文即将完成之际，我在这里向所有帮助过我的人致以最诚挚的谢意。

首先，我要由衷的感谢我的研究生导师杨震教授。杨老师严谨的学术态度和悉心的教导深深的影响了我，指引我前行，无论是学术上还是生活中都给予了我莫大的帮助。杨老师虽然工作繁忙，但从不落下对我们的指导和帮助，百忙之中也不忘关注我们的科研进展，及时给予指导，我们的大小论文他都会仔细修改，甚至连错别字都会一一找出。我这篇论文的完成离不开杨老师的谆谆教诲，再次感谢杨老师。同时还要感谢崔景伍老师，感谢她对我们在生活上的支持和关怀，为我们处理各类财务和报销问题。

每次回想起来，我都会非常庆幸加入了杨门大家庭，这里的同门兄弟姐妹们都像家人一般，在我遇到困难时总会伸出援手，为我排忧解难。感谢吕斌、郭海燕、田峰和杨真真老师在日常的科研中对我的指导和提出的宝贵意见；感谢陈海霞师姐在我对自己的研究领域一无所知时为我解答疑惑；感谢同届的祖婉婉、时安谊同学在平常的生活中互相关心和照顾；感谢师弟师妹们的帮助。再次感谢曾经帮助过我的所有人。

最后，非常感谢各位专家和教授能够抽出宝贵的时间审阅我的论文、参加我的论文答辩，衷心感谢各位的批评指导。