**清 华 大 学**

**综 合 论 文 训 练**

　题目：　深度声学特征学习及其应用

系　　别：　自动化系

专　　业：　自动化

姓　　名：　王赫麟

指导教师：　高飞飞 副教授

2019年6月10日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 导师签名： 日 期：

中文摘要

近年来，随着深度学习的发展，基于ImageNet数据库的视觉目标识别取得了巨大突破。其中，基于深度神经网络(如VGGNet)所训练的深度图像特征模型，成为了通用特征提取方法并广泛应用于其它机器视觉任务。然而，目前针对音频处理任务，尚无类似的深度声学特征模型。传统的人工声学特征提取方法往往针对任务而设计，而对于其他的声学处理任务缺乏有效性。

本文采用一定规模的音频数据库，进行深度声学特征模型的设计和学习，并测试了基于此模型的多任务系统的性能。本文提出一种基于CNN的深度声学特征提取网络，称为SCNN网络。第一层使用SincConv卷积层，具有较好的解释性，收敛速度快。同时，本文介绍了基于SCNN网络的声学事件检测系统、声学场景分类系统及说话人识别系统，均为端到端系统。通过在不同的数据集上进行模型训练和测试，均实现了较好的效果。实验证明，文中提出的方法可以在多个声学任务上有效提取声学特征。

关键词：深度学习；深度声学特征；声学事件检测；声学场景分类；说话人识别

ABSTRACT

In recent years, with the development of deep learning, great breakthroughs have been made in visual object recognition based on ImageNet database. Among them, deep image feature model trained by deep neural network (such as VGG) has become a general feature extraction method and is widely used in other machine vision tasks. However, there is currently no similar deep acoustic feature model for audio processing tasks. Traditional methods of hand-crafted acoustic feature extraction are often designed for specific tasks, but lack of effectiveness for other acoustic processing tasks.

In this paper, I use some medium-sized audio datasets to design and learn the deep acoustic feature model, and then the performance of the multitasking system based on this model is tested. This paper proposes a CNN-based deep acoustic feature extraction network called SCNN network. The first layer uses the SincConv convolutional layer, which is better interpreted and converges faster. At the same time, this paper introduces the acoustic event detection system, acoustic scene classification system and speaker recognition system based on SCNN network, which are end-to-end systems. Good results were achieved by performing model training and testing on different datasets. Experiments show that the proposed method can effectively extract acoustic features on multiple acoustic tasks.

**Keywords:** Deep learning; Deep acoustic feature; Acoustic event detection; Acoustic scene classification; Speaker recognition

目 录

[第1章 引言 1](#_Toc10025948)

[1.1 课题研究的背景与意义 1](#_Toc10025949)

[1.2 课题的研究现状 1](#_Toc10025950)

[1.3 研究内容和论文组织结构安排 2](#_Toc10025951)

[第2章 声学特征研究概述 4](#_Toc10025952)

[2.1 声学特征提取方法概述 4](#_Toc10025953)

[2.2 信号表示 4](#_Toc10025954)

[2.2.1 信号采集和预处理 5](#_Toc10025955)

[2.2.2 通用时频表示 5](#_Toc10025956)

[2.2.3 对数频率和感知机理表示 7](#_Toc10025957)

[2.3 特征设计 8](#_Toc10025958)

[2.3.1 时域特征 8](#_Toc10025959)

[2.3.2 频域特征 9](#_Toc10025960)

[2.3.3 倒谱特征 9](#_Toc10025961)

[2.3.4 感知机理特征 10](#_Toc10025962)

[2.3.5 图像特征 11](#_Toc10025963)

[2.4 特征学习 12](#_Toc10025964)

[2.4.1 基于CNN的深度声学特征提取 12](#_Toc10025965)

[2.4.2 基于DNN的深度声学特征提取 13](#_Toc10025966)

[2.5 降维和特征选择 13](#_Toc10025967)

[2.5.1 降维 13](#_Toc10025968)

[2.5.2 特征选择 13](#_Toc10025969)

[2.6 时域积分和池化 14](#_Toc10025970)

[2.6.1 基于统计的时域积分 14](#_Toc10025971)

[2.6.2 基于模型的时域积分 14](#_Toc10025972)

[2.7 本章小结 15](#_Toc10025973)

[第3章 基于CNN的深度声学特征模型设计 16](#_Toc10025974)

[3.1 模型设计思路 16](#_Toc10025975)

[3.1.1 方案及对比 16](#_Toc10025976)

[3.1.2 模型设计概述 17](#_Toc10025977)

[3.2 SincConv卷积层概述 18](#_Toc10025978)

[3.3 后端CNN网络设计概述 20](#_Toc10025979)

[3.4 相关方法介绍 26](#_Toc10025980)

[3.4.1 激活函数 26](#_Toc10025981)

[3.4.2 损失函数 26](#_Toc10025982)

[3.4.3 优化算法 26](#_Toc10025983)

[3.4.4 梯度问题和过拟合问题的处理 27](#_Toc10025984)

[3.5 本章小结 27](#_Toc10025985)

[第4章 基于SCNN网络的声学事件检测和声学场景分类系统 28](#_Toc10025986)

[4.1 引言 28](#_Toc10025987)

[4.2 声学事件检测和声学场景分类任务概述 28](#_Toc10025988)

[4.3 基于SCNN网络的声学事件检测系统概述 29](#_Toc10025989)

[4.4 基于SCNN网络的声学场景分类系统概述 31](#_Toc10025990)

[4.5 实验设置及结果分析对比 31](#_Toc10025991)

[4.5.1 基于SCNN网络的声学事件检测系统实验设置 31](#_Toc10025992)

[4.5.2 基于SCNN网络的声学场景分类系统实验设置 32](#_Toc10025993)

[4.5.3 实验结果分析及对比 33](#_Toc10025994)

[4.6 本章小结 37](#_Toc10025995)

[第5章 基于SCNN网络的说话人识别系统 39](#_Toc10025996)

[5.1 引言 39](#_Toc10025997)

[5.2 说话人识别任务概述 39](#_Toc10025998)

[5.3 基于SCNN网络的说话人识别系统概述 40](#_Toc10025999)

[5.4 实验设置及结果分析对比 40](#_Toc10026000)

[5.4.1 实验设置 40](#_Toc10026001)

[5.4.2 实验结果分析及对比 41](#_Toc10026002)

[5.5 本章小结 42](#_Toc10026003)

[第6章 总结与展望 43](#_Toc10026004)

[6.1 论文工作总结 43](#_Toc10026005)

[6.2 未来工作展望 43](#_Toc10026006)

[插图索引 45](#_Toc10026007)

[表格索引 47](#_Toc10026008)

[参考文献 48](#_Toc10026009)

[致谢 51](#_Toc10026010)

[声明 52](#_Toc10026011)

[附录A 外文资料的中文翻译 53](#_Toc10026012)

1. 引言
   1. 课题研究的背景与意义

人工智能[6](Artificial Intelligence, AI)是计算机科学的一个重要领域，旨在实现类似人类那样思考、甚至超过人类的智能体。从声学的角度讲，实现人工智能就是让机器像人一样具有听、说能力，并且会思考。如果机器可以实现和人类一样的听觉能力，我们希望它能够区分各种声音，提取出人们感兴趣的部分进行处理。

由于音频输入和应用场景的多样性，机器听觉[4]是一项极其复杂和艰巨的任务。通常，人们会将其分为一些较小的子问题，并且大多数的研究工作都集中在解决更加简单、具体的任务上。实际上，设计能够处理不同类型声音的通用机器听觉系统是一项真正有挑战性的工作。相反，由于系统设计可以被调整和优化以考虑其信号特性，开发能够完成特定任务和限于特定性质的声信号的系统往往更加容易。

目前为止，语音[1][7]和音乐[8]是机器听觉背景下最广泛研究的声源类型，呈现出一些独特的特征。相反，来自环境的其它类型的声源（例如，交通噪声，来自自然界中的动物的声音等）没有表现出这种特性，或者至少没有表现出这种特殊性。尽管如此，这些非语音和音乐相关的声音也可以作为声学事件[9]或声学场景[10]来检测和识别。无论其具体目标如何，任何机器听觉系统都需要对输入的音频信号进行深入分析，目的是为了充分利用其特定的属性。这个过程通常被称为声学特征提取。

传统研究表明[5]，声学信号的特征主要有时域和频域两种，并有很多提取这两种重要特征的方法。时域特征包括短时平均能量、短时平均过零率、共振峰、基因周期等；频域特征有线性预测系数、LP倒谱系数、线谱对系数、短时频谱、Mel频率倒谱系数等。还有结合时间和频率的特征，如频谱图。这些特征通常被称为人工声学特征。一般来说，这些人工声学特征是根据特定需求而设计，只包含了声学信号的部分信息。然而，为了满足具体的声学处理任务，人工声学特征往往要经过变换、取舍与组合等处理。因此，如何更加充分地提取信息，不囿于所研究问题而得到相对通用的声学特征，就成了一个有意义和有挑战性的问题。

* 1. 课题的研究现状

通常来说，声音信号或波形的时域表示不容易直接解释声音信息。大多数情况下，仅仅依靠波形几乎不可能完成音频的识别或分类等任务。因此，频域表示和时频域表示（包括多尺度表示）已经被人们使用多年，以提供更符合人类感知的声音信号的表示。

然而，这些表示通常过于通用，并且经常无法描述音频中存在的特定信息。许多工作都致力于设计可以提取这些特定信息的特征，从而产生各种人工声学特征提取方法。这些特征的一个问题是，根据设计，它们可以满足特定任务的要求，但是通常不能很好地适用于其他任务，鲁棒性较差。它们通常需要与其他特征组合，从而产生较大的特征向量，这个过程被称为特征设计[11]。但在过去几年中，由于中等规模和大规模声音数据集的可用性日益增加，特征设计的替代方法开始变得流行，所谓的特征学习已被证明与大多数精细调整的人工特征相比更具有竞争力。实际上，越来越多的大规模的数据集已经可用于开发声学特征学习技术。非负矩阵分解，稀疏表示学习，字典学习和深度学习的发展是这种趋势的表现。这些方法能够提取反映特定任务中考虑的数据的基础结构的特征，提供可以在某种程度上概括在训练阶段期间未见的数据构造的高级表示。2006年，随着深度置信网络(Deep Belief Networks, DBN)的提出，神经网络的研究开始兴起[3]。深度神经网络(Deep Neural Networks, DNN)即可由训练好的DBN与一个基于任务的输出层组成。同时，由于大数据时代的到来，以及GPU等计算资源的出现，深度神经网络的训练成为可能。深度神经网络已经在许多领域发挥着重要作用，包括计算机视觉、语音识别和自然语言处理等。采用深度神经网络进行声学特征学习也成为主流研究。

同时，无论采用传统的特征设计方法还是特征学习方法，对于越来越庞大的数据集来说都显得难以应对。因此，能够通过减少特征向量维度或通过减少要处理的特征向量的数量来减小特征空间中的数据集的大小是至关重要的。

* 1. 研究内容和论文组织结构安排

声学特征提取是几乎所有声学处理任务的前提，同时所提取的声学特征也往往是决定声学处理系统性能的重要影响部分。目前，随着深度学习的发展，特征学习已成为越来越受欢迎的特征提取方法。主流的研究仍聚焦于通过深度学习的方法完成特定任务的声学特征学习，本文提出一种基于CNN的深度声学特征模型，用以提取原始波形的深度声学信息，同时可以在声学事件检测、声学场景分析以及说话人识别等多个任务上有较好的表现。

本文共有六个章节，具体每一章的内容安排如下：

第一章绪论：展示了本文的研究背景，传统的声学特征提取方法针对特定任务进行设计，往往无法在其他任务上取得好的效果，随着深度学习的快速发展，采用深度神经网络进行特征学习使得在多任务上实现特征提取成为可能。在本章，作者详细叙述了本文的主要研究内容、课题研究的背景及意义、研究现状和论文结构安排。

第二章声学特征研究概述：本章首先展示了声学特征提取的流程，接着介绍了传统的人工声学特征提取方法，然后介绍了基于深度神经网络的声学特征提取方法，为下文介绍本文设计的网络模型和系统结构做铺垫。

第三章基于CNN的深度声学特征模型设计：阐述了模型设计思路，提出一种基于CNN的深度声学特征模型，称之为SCNN网络。紧接着对于第一层SincConv卷积层和后端CNN网络设计进行了详细的介绍，然后介绍了模型中相关方法的使用，包括激活函数的选择、梯度问题的处理、损失函数的选择以及优化算法，该模型是下文各个系统的前端结构。

第四章基于SCNN网络的声学事件检测和声学场景分类系统：本章首先介绍了声学事件检测和声学场景分类任务，然后介绍了基于SCNN网络的声学事件检测系统和声学场景分类系统结构，通过实验对二者的性能进行评估，这里详细介绍了实验配置情况，最后对实验结果进行了分析，并将其与标准CNN网络结构进行对比分析。

第五章基于SCNN网络的说话人识别系统：本章将SCNN网络用于说话人识别任务进行特征提取。首先对说话人识别任务进行概述，接着介绍基于SCNN网络的说话人识别系统结构，同时通过实验对其性能进行评估。

第六章总结与展望：对论文工作进行总结，对本文提出的方法及实验结果进行总结，指出本文工作的创新点和不足之处，最后对未来的发展和方向进行展望。

1. 声学特征研究概述
   1. 声学特征提取方法概述

本章介绍了将声音信号转化为可被分类器等后端系统有效利用的特征向量的通用处理链，这个过程被称为声音信号的特征提取[12][2]。标准步骤如下图 2.1所示，首先需要从声音信号中提取有用信息的适当表示(详见2.2)，然后采用特征设计(详见2.3)或者特征学习(详见2.4)的方法提取特征，接着对提取到的特征进行选择(详见2.5.2)和降维处理(详见2.5.1)，最后，对于一些需要考虑时间信息的任务，往往还需要时域积分(详见2.6)。

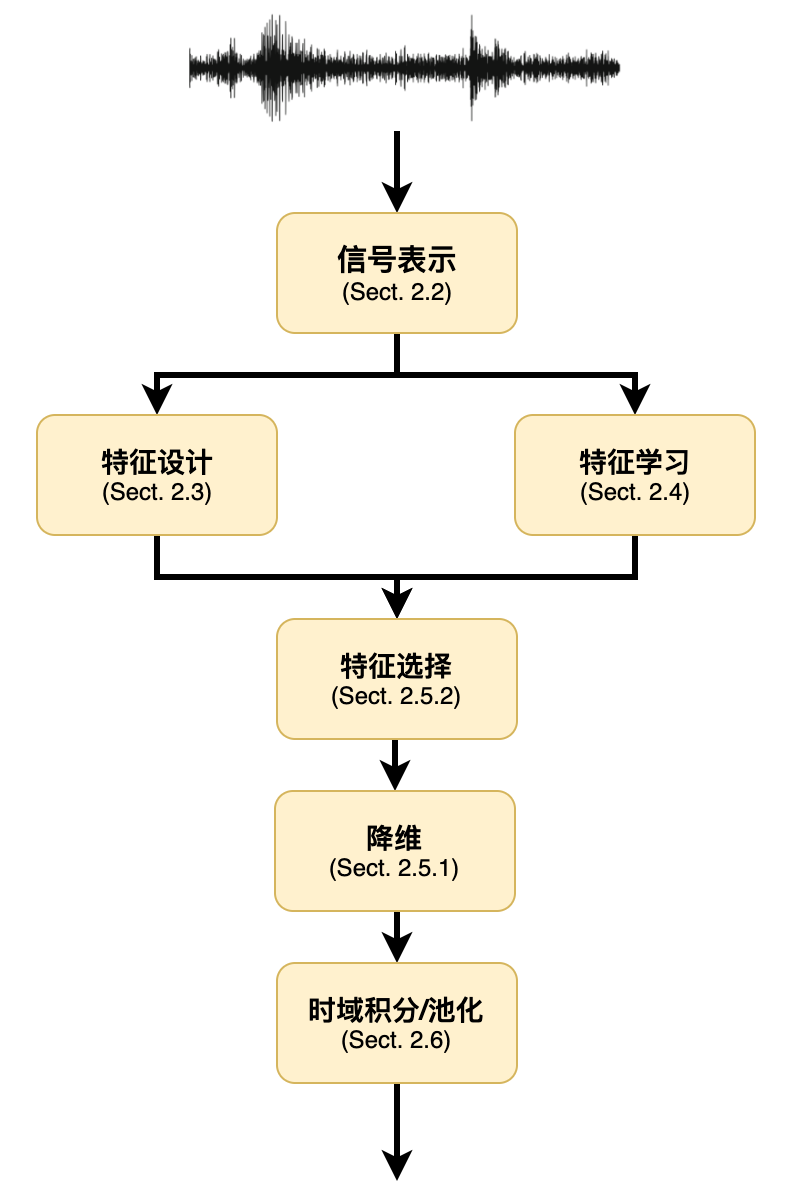


图 2.1声学特征提取标准流程图

* 1. 信号表示

许多工作都致力于找到从声音信号中提取有用信息的适当表示。本节介绍了一些主要的声音信号表示。

* + 1. 信号采集和预处理

一般而言，声音是振动通过诸如空气或水之类的介质传播的结果。可以通过麦克风等的电声转化器以电信号的形式记录声音，然后可以将该模拟信号转换为数字信号,并在进一步分析之前存储在计算机上。执行此模数转换的必要步骤包括：

1. 滤波阶段：对模拟信号进行低通滤波，使得将其频率带宽限制在区段内，其中是低通滤波器的截止频率。

2. 采样阶段：然后以采样速率对低通模拟信号进行数字采样，以避免频率混叠现象。

3. 量化阶段：然后量化所获得的数字信号（例如信号的幅度只能采用有限数量的预定义值来保存存储容量。）

4. 可选的附加阶段：在某些情况下，可以执行其他预处理阶段，例如预加重。该步骤可以在简单的一阶有限脉冲响应(FIR)高通滤波器的形式下执行。

音频CD质量的典型值是的采样率和每个样本的量化，故单通道音频信号的比特率为。更高的质量标准包括,或的采样率和的量化。

* + 1. 通用时频表示

在进行任何分析之前，声音信号通常会转换为频域。利用离散时间傅里叶变换(DFT)可以获得线性频率尺度上的信号的频域表示，如所示。

其中为得到的频谱，频率 表示奈奎斯特频率， 表示采样频率。可以使用逆离散时间傅立叶变换(IDFT)将频谱变换回时域，如所示。

实际上，常常通过在信号的长度为N的窗口帧上使用DFT来近似得到频谱。这被称为短时傅立叶变换(STFT)。的第帧的DFT的第个分量计算，如所示。

式中，是窗函数(如矩形窗，汉明窗，布莱克曼窗等)，用来衰减由DFT近似带来的一些影响和保证帧边缘的连续性和周期性。帧之间的跳跃等于帧的长度(N)，这意味着连续帧之间没有重叠。通常选择小于帧长度的跳跃大小以便完整保留音频信号。

时域信号的第帧可以使用逆STFT从离散频谱中获得。而STFT和逆STFT都可以通过快速傅立叶变换(FFT)和逆快速傅立叶变换(IFFT)计算。

STFT定义了线性频谱图，是声音的二维表示，其中每个频带中的能量作为时间的函数给出。该频谱图是矩阵，其中每列是声音信号帧的DFT的模数(见图 2.2b)。

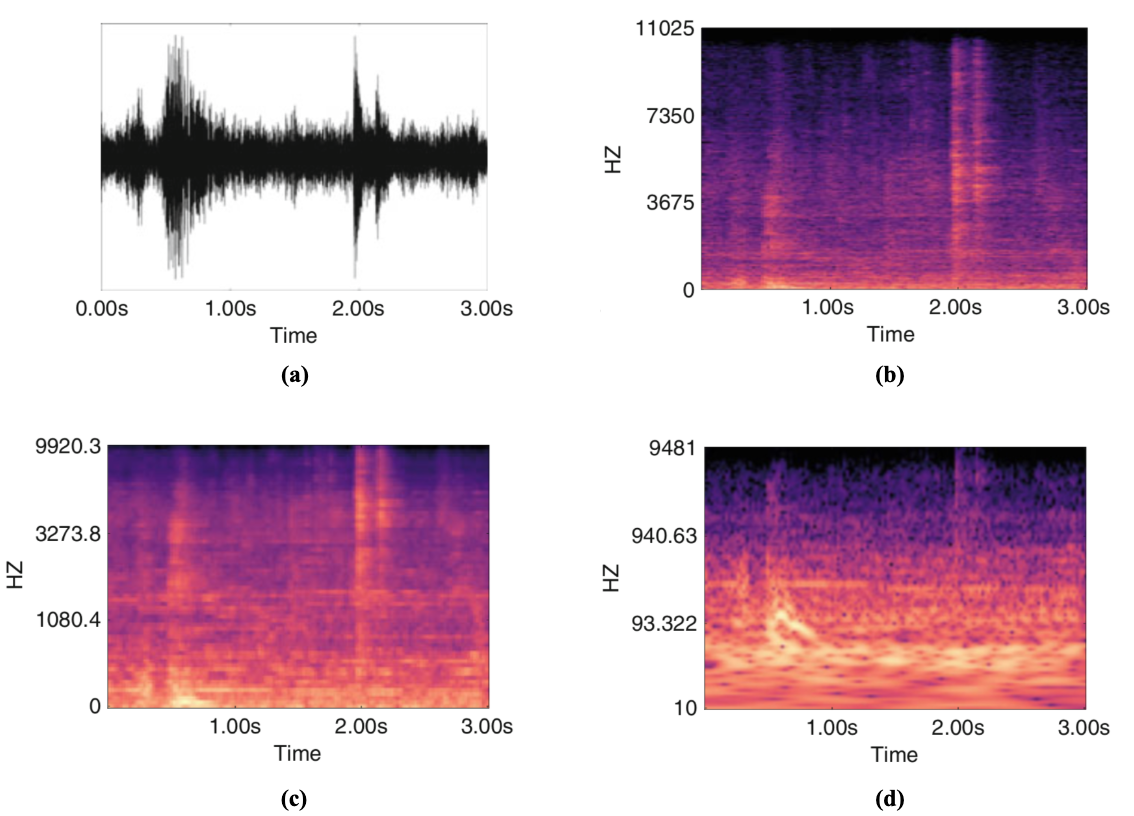


图 2.2在餐厅录制的声音信号的不同时域和时频域表示图：0.5秒时有人正在清理喉咙，2秒时有一些餐具噪音[13]。(a)时域波形。(b)线性频谱图。(c)Mel谱图。(d)Constant-Q谱图。

* + 1. 对数频率和感知机理表示

人们通常希望得到特定频带中的信息，这可以通过计算预定义频带中的能量或能量比来实现(见图 2.2c)。频带可以在频率轴上等间隔，根据对数或感知定律来设置。频带的数量，原型滤波器的形状以及频带之间的重叠也可以有很大变化。

1. 关键频带是由Fletcher提出的[14]。其核心想法是关键频段描述了耳蜗中听觉滤波器的带宽。从概念上讲，这意味着同一临界频带内的两个音调将相互干扰，这就是所谓的频率屏蔽现象。等效矩形带宽标度(ERB)提供了一种计算近似于听觉滤波器的矩形滤波器的中心频率和带宽的方法[15]，如式所示。

式中，的单位是Hz。

2. Gammatone过滤器是线性滤波器，其脉冲响应是由正弦波载波(音调)组成，其幅度由一个包络调制，该包络具有与缩放的伽玛分布函数相同的形式，如式所示。

式中，是幅度，是滤波器阶数，是时间衰减系数(与滤波器带宽有关)，是载波的频率(与滤波器的中心频率有关)，是载波的相位(与包络位置有关)。与ERB类似，已证明4阶的Gammatone滤波器可以很好地逼近听觉滤波器[16]。

3. Mel 标度对应于纯音音高(例如纯正弦曲线)的心理感觉的近似情况[17]，一种Mel标度和Hertz标度之间常用的关系，如所示。

4. Constant-Q变换(CQT)与DFT密切相关。一个主要的区别是，不是使用频率之间具有恒定间隔的频率标度(如在DFT中)，而频率是几何分布的[18]。这产生频带的中心频率与频率分辨率之间的恒定比率。第个频段的中心频率由式给出：

其中，表示第一频段的中心频率，表示每个八度音程的频率数(见图2d)。CQT最初是为了映射西方音乐标度而引入的。

* 1. 特征设计

特征设计的方法广泛应用于各类音频处理任务，是一种精细制作来满足低级别表示的特殊功能的方法，依赖于类不变性的知识。本节介绍了一些常见的特征类型。(见图 2.3)

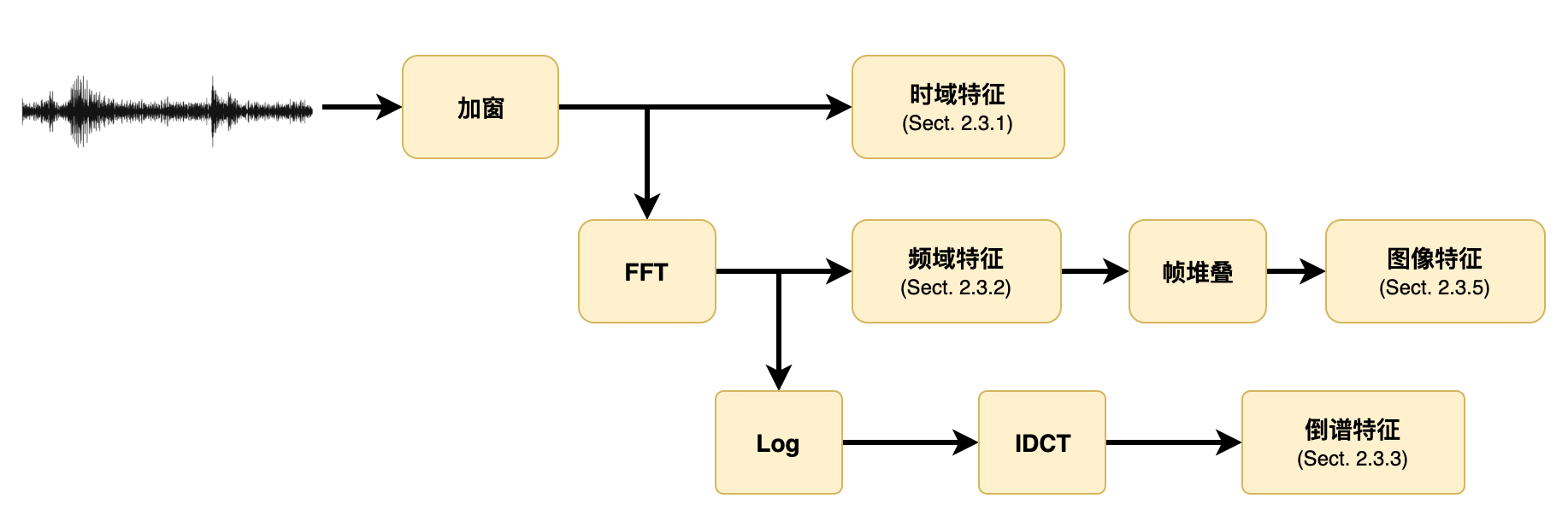


图 2.3特征设计过程示意图

* + 1. 时域特征

时域特征直接在时间波形上计算，因此通常很容易计算。下面描述了一些最常见的时间特征。

1. 时域包络可以被视为包含信号的边界, 它是静音检测的可靠指标。一个简单的实现依赖于计算信号的平均能量的均方根,其中大小为的帧跨越时间索引：

2. 过零率(ZCR)由信号幅度与零值相交的时间给出。对于大小为的帧，它由式给出：

其中，返回信号幅度的符号。该特征使用广泛，因为它可以很简单地将周期性信号(过零率小)与被一定程度随机噪声(过零率大)破坏的信号区分开。

3. 自相关系数可以被解释为时域中的信号频谱分布。实际上，通常只考虑可以获得的前个系数，如式所示。

* + 1. 频域特征

对于声音感知的研究广泛依赖于声音信号的频率成分。下面介绍几种常见的频域特征。

1. 能量是最直接但最重要的频谱特征之一。该特征可以直接计算为频段中振幅分量的平方的总和。计算频段中的对数能量也很常见。

2. 频谱包络在概念上类似于时域包络，只是在频域中。它可以被视为包含信号频谱的边界。例如，可以使用线性预测编码（LPC）来近似频谱包络[19]。

3. 频谱矩描述了一些主要的频谱形状特征，包括频谱形心，频谱宽度，频谱不对称性和频谱平坦度。

4. 幅度谱平坦度是频谱平坦度特征的替代方案，可以通过频谱幅度的几何和算术平均值之间的比率（全局或几个频带）来计算。

5. 频谱斜率测量频谱随频率降低的平均速率。

6. 频谱滚降定义为存在总频谱能量的预定百分比（通常在85％和99％之间）的频率。

7. 频谱通量表征频谱信息的动态变化，可通过幅度谱的导数或者作为连续幅度谱之间的归一化相关性来计算。

8. 频谱不规则特征旨在得到与声音部分（例如声音的各个频率分量）相关联的更精细的信息描述。

在很多声学任务中，时域特征和频域特征很少单独使用。事实上，它们大多是简单的功能，旨在模拟信号的特定方面，因此通常需要与其他特征相结合。 对数梅尔能量特征是一个例外，可以单独用作分类或特征学习的输入。

* + 1. 倒谱特征

倒谱特征是根据广泛用于模拟语音产生的所谓源滤波器模型来分解信号，然后将信号分解成载波（源，对于语音来说是声门激发）和调制（滤波器，对于语音来说包括声道和舌头的位置）。

1. 梅尔频率倒谱系数(MFCC)是最常用的倒谱系数[20]。它们是在梅尔频带中对数能量的逆离散余弦变换获得的，如式所示。

这里的是梅尔频带数目，是梅尔频带索引，是第个梅尔频带的能量，是倒谱系数的索引()。实际上，一种常见的实现方式是使用三角形滤波器组，其中每个滤波器根据梅尔频率标度进行间隔（见图 2.4a）。获得频带中的能量系数作为频谱幅度分量的加权和。（根据相应三角形滤波器的幅度值给出权重）。对于16kHz的带宽，滤波器的数量通常在12和30之间变化。MFCC广泛用于语音处理，同时也是声音场景分析最受欢迎的特征之一。

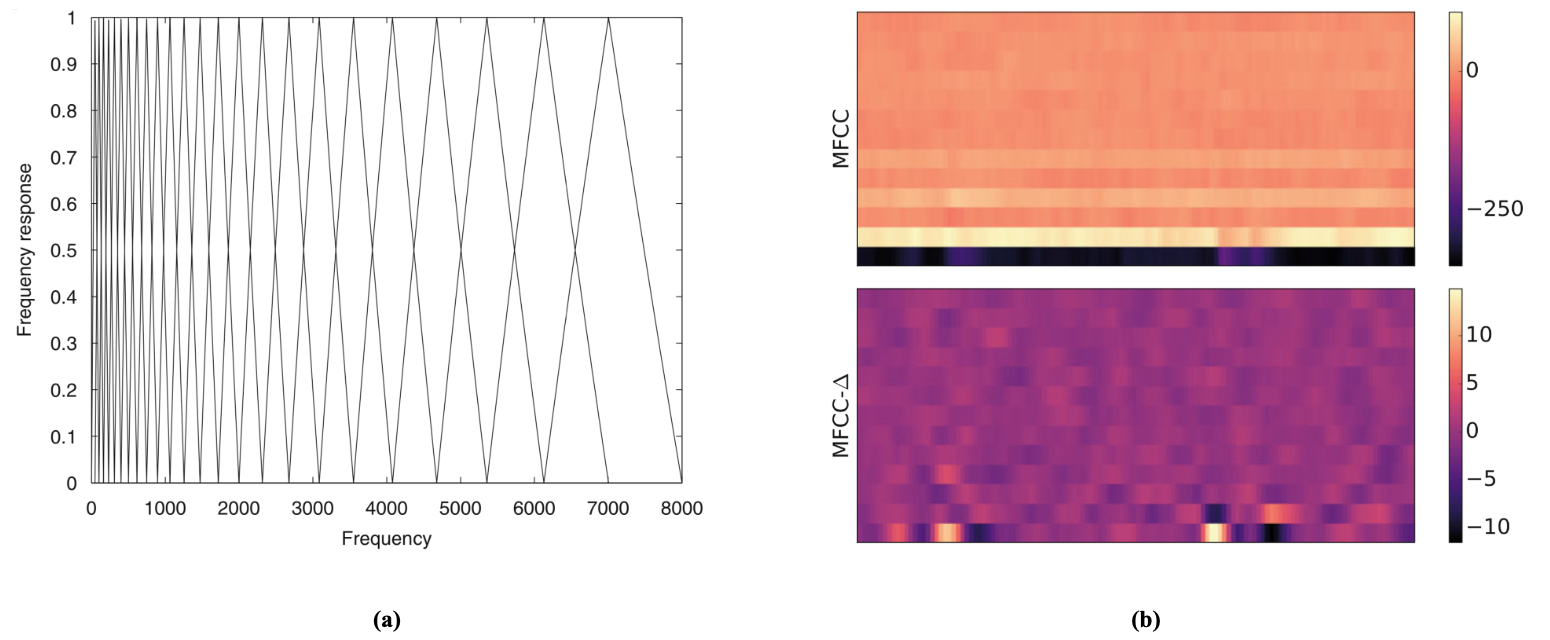


图 2.4梅尔滤波器频带(a) MFCC分解(b)

2. 与MFCC类似，可以通过其他频域表示获得一些替代的倒谱分解，比如基于LPC系数的线性预测倒谱系数(LPCC)，Gammatone特征倒谱系数(GFCC)或恒定Q倒谱系数(CQCC)等特征[21]。

* + 1. 感知机理特征

对人类感知的研究使人们能够更好地理解人类听觉过程。这些研究的一些结果（例如听觉滤波器）已在特征设计中得到利用，并产生广泛使用的特征，如MFCC。在特征提取中可以利用各种各样的感知属性。下面描述了三种感知功能：

1. 响度（以声音测量）是声音强度的主观印象，使得声音加倍对应于响度的加倍。它通常由特定响度在所有ERB频段上的积分获得，如式所示。

式中，是ERB频段的数目，每一个频段上的响度可以由式近似得到：

其中，是第个频带的信号的能量。

2. 清晰度可以解释为基于心理声学原理的频谱形心，通常被估计为特定响度的加权形心。

3. 感知扩散是给定声音的音宽的度量，通过最大特定响度和总响度之间的相对差异计算，如式所示。

* + 1. 图像特征

基于频谱图像的特征是从声音的时频表示中提取的特征。受计算机视觉的启发，图像特征旨在表征时频内容的形状，纹理和演变。

1. 定向梯度直方图(HOG)是计算机视觉中用于在图像中执行形状检测的基于图像的特征。它是根据声音的频谱图像计算出来的，目的是捕捉相关的时频结构来表征声音信息[22]。通常通过计算包含频谱图像中每个像素的梯度图像来提取。梯度图像的每个像素表示原始图像中强度变化的方向。在非重叠单元中分离图像之后，在每个单元中计算每个像素的梯度方向的直方图。

2. 子带功率分布(SPD)是将时频图像转换为频率与频谱功率的二维表示。它是通过估计谱图的每个子带中的谱分布来计算的。实际操作中，通过提取每个子带中的像素值的直方图来估计分布。SPD图像可以直接用作特征，也可以用作提取其他基于图像的特征的中间表示。

3. 局部二值模式(LBP)分析是一种用于图像识别的特征提取技术，用于表征图像中的纹理。LBP特征是与图像中的每个像素相关联的二元矢量。它们是通过将给定像素的值与固定邻域中的其他像素进行比较而构建的。LBP已应用声场分析，以捕捉场景谱图的纹理和几何属性。

* 1. 特征学习

最近研究表示，特征学习技术优于许多分类和其他声音分析任务中的人工设计特征。实际上，越来越多的较大规模的数据集已经可用于开发特征学习技术。非负矩阵分解，稀疏表示学习，字典学习和深度学习的发展是这种趋势的表现。本节主要介绍基于深度学习的特征学习技术。

在过去的十几年中，训练算法和计算能力方面的进步导致了深度学习技术的普遍使用，这些技术现在已成为许多音频应用中的最新技术。在模式分类中最常见应用的深度学习技术，如深度神经网络（DNN），卷积神经网络（CNN），递归神经网络（RNN）均可用于学习特征。下面简要介绍采用CNN和DNN的方法进行深度声学特征提取。

* + 1. 基于CNN的深度声学特征提取

自从2010年ImageNet[23]国际计算机视觉挑战赛(ILSVRC)创办以来，涌现很多经典的网络结构模型：AlexNet[25]、VGGNet[24]、Google Inception Net[26]和ResNet[27]等。这些网络结构模型均采用CNN网络，在特征提取上效果显著。尽管MFCC等标准的人工声学特征仍然应用于许多先进的系统，但由于CNN独特的卷积计算，神经元之间采用局部连接而且一组连接共享同一个权重，比简单的投影、方向和重心等更加可靠，整体特征模型的泛化能力更强，向CNN馈送频谱图像或甚至直接利用原始音频信号成为越来越受欢迎的方法。通过向二维CNN馈送频谱图像的方法，属于二维声学特征提取[31]；而将原始音频信号利用一维CNN进行时序分析的方法，属于一维声学特征提取[28][29]。

二维声学特征提取主要借鉴了计算机视觉领域的特征提取方法。首先，提取音频信号的频谱图或Log-Mel能量，作为图像输入CNN网络中进行特征的提取。这种方法使得音频的时频特征得到较好的表示。

手工特征最初是根据人类感知进行设计的，不能保证这种表示对于所有的声学任务是最佳的。例如，标准特征使语音频谱平滑，可能妨碍提取关键的窄带扬声器特征，如音调和共振峰。相反，直接处理原始波形使得网络在每个特定任务上可能学习到更加适用的低级表示。一维声学特征以此为出发点，直接对原始波形进行时序建模，通过卷积计算提取深度特征。与二维声学特征提取方法相比，这种方法不受制于频谱图的提取方式的影响，更加充分地利用了音频信号中的信息。

* + 1. 基于DNN的深度声学特征提取

一些研究首先提取手工特征，然后在手工特征的基础上进一步提取深度特征，一般会采用DNN网络的方式。比如，一种使用多流分层DNN提取深度音频特征(DAF)的方法[30]。该方法结合MFCC和Gabor特征，训练每个隐藏层以提取于区分音频信息，实验验证效果优于手工特征。

* 1. 降维和特征选择

在音频处理系统的设计中可以考虑许多潜在有用的特征。尽管将所有这些特征结合处理任务有时是可行的，但这样做可能是次优的，因为它们中的许多可能是冗余的，甚至会由于非鲁棒的提取过程而产生噪声。因此，特征选择或压缩变得不可避免，以便通过减少其维度来降低问题的复杂性，并且仅保留与任务相关的信息。

* + 1. 降维

处理特征空间的大维度的常用方法是PCA，线性判别分析（LDA）等变换技术，更新的方法是设置瓶颈层DNN。这里不做详细叙述。

* + 1. 特征选择

特征选择是PCA等特征变换技术的替代方式，主要是考虑到PCA等需要在测试阶段提取所有候选特征的不便。此外，PCA不保证能够消除噪声特征，并且变换的特征难以解释。特征选择旨在解决这些问题。

通过特征选择(FS)，从大的个候选集合中选择个特征的子集，目的是实现任务最低的Loss。这项任务非常复杂：由于涉及到极多的组合，当预先没有给出时搜索空间的大小是,因此执行详尽的子集搜索是不切实际的，而且计算每个候选特征子集的Loss成本也很高。因此，通常通过引入两个主要简化方式来以次优方式解决特征选择：

1. 通过重复使用接近最优的搜索策略，可以避免暴力搜索。

2. 优先选择更简单的特征选择标准，而不是使用Loss评价，作为预处理阶段的一部分，在本质上是利用初始特征集。

* 1. 时域积分和池化

上述大多数功能通过短时信号分析窗口（或帧）来捕获给定信号的特定属性，在此窗口上信号可被视为静止。通常假设不同帧中的特征的连续观察在统计上是独立的，这意味着会忽略掉这些特征的时间演变。本节描述了几种方法，通常称为时域积分，以考虑信号时间演化中传达的信息。

* + 1. 基于统计的时域积分

可以在短分析帧上局部计算的“瞬时”特征上直接进行时间积分，这种前项积分通常在较大的时间窗口上完成。前项时间积分过程可以用函数表示，函数应用于特征向量序列，记为,其中对应于在第帧观察到的第个标量特征。

积分函数的目的是捕获短时统计(例如下面描述的均值和协方差)或使用某种模型进行更复杂的时间积分。前项积分是计算特征过程的一阶统计量。然后将均值积分函数定义为：

这种简单的方法可以扩展到最大值和绝对值池化或更高阶统计量，例如使用完全协方差矩阵，偏度或峰度等。

* + 1. 基于模型的时域积分

更复杂的模型也可用于模拟连续特征之间的时间依赖性。例如，可以将特征序列建模为自回归过程。这样的模型将捕获一些全局频谱特性，其中细节水平取决于AR模型的顺序。按照Meng[47]中用于音乐类型分类的多变量自回归模型，相应的积分函数可以写成：

其中，和是第个窗口的模型参数的最小二乘估计，并且由表示的阶模型定义为：

这里是D维白噪声向量。

* 1. 本章小结

本章概述了标准特征提取过程的通用处理链。大多数特征提取技术的第一步是选择适合的时频表示。作为处理链的起点，它们在构建音频任务系统中起着至关重要的作用。

本章还介绍了一系列最常用的人工特征。然而，它们通常仅限于描述时频信息的特定部分。人工特征的功能受到限制，因为通常需要结合各种不同的人工特征来提高性能。

最后，通过使用深度神经网络技术的特征学习，是一种越来越受欢迎的特征提取方法，具有可以直接从时频表示自动学习数据中的相关信息的优点。因此，往往可以省略掉特征选择和降维等操作。但是，深度学习技术需要一定的训练过程才能找到目标任务的适当架构，这很大程度上取决于数据集的内容和大小。可以预见，该领域的未来发展会更加重视新的更大的数据集，这将进一步提高深度学习技术的有效性。

1. 基于CNN的深度声学特征模型设计
   1. 模型设计思路
      1. 方案及对比

考虑到特征设计的局限性以及深度学习技术下特征学习的优势，本文提出一种基于深度神经网络的声学特征提取方法。下表 3.1通过对比几种主流的方法，确定本文的研究方案。

表 3.1 几种主流深度特征提取方案的对比

|  |  |  |  |
| --- | --- | --- | --- |
|  | 方案一 | 方案二 | 方案三 |
| 名称 | 利用声谱图提取深度特征 | 利用MFCC等特征提取深度特征 | 利用原始音频提取深度特征 |
| 描述 | 前向为传统的声谱图，借鉴图像特征提取方式，通过CNN网络进行深度特征提取。 | 前向为MFCC,PLP等手工特征，设计DNN网络进一步提取其深度特征。 | 直接输入原始音频信号，通过设计深度网络模型（CNN等）提取其深度特征。 |
| 优点 | 思路简单，  容易实现。 | 较可靠，  较容易实现。 | 可以保留完整的音频信息，可靠，属于主流设计方案。 |
| 缺点 | 采用使用短时傅里叶变换后的声谱图丢失相位信息，仅保留幅度信息，得到的深度特征不够完全，效果可能不佳。 | 虽然MFCC等特征已被广泛应用，但仍存在信息不全、易受噪声影响等缺点。同时，采用不同的特征结合方式，差异较大，设计合适的网络较难。 | 网络设计较难，小规模数据集效果可能不佳。 |

考虑到上述方案的优缺点，本文提出一种直接向深度神经网络馈送原始音频信号来提取声学特征的方法，这种方法属于端到端的方法，抛弃了手工声学特征，对于音频信息的保留度高，更有可能实现多任务下的声学特征提取。

* + 1. 模型设计概述

本文采用基于CNN的一维声学特征提取方法，输入为原始波形信号，网络结构借鉴了VGGNet[24]，第一层采用SincConv卷积层[29]取代标准卷积层结构，下图 3.1是基于分类任务下的网络结构图。卷积层之间采用Batch Normalization[32]进行批标准化，激活函数采用Leaky ReLU函数，并且会进行最大池化处理来减少特征参数，增加网络的泛化能力，层间结构见图 3.2。

整体网络模型称之为SCNN(Convolution Neural Network with SincConv Layer)。本章详细介绍了网络结构的细节，3.2节介绍了SincConv卷积层，3.3节介绍了后端卷积层的设计，3.4节介绍了相关方法，包括激活函数和损失函数的选择、梯度问题和过拟合问题的处理、优化算法。

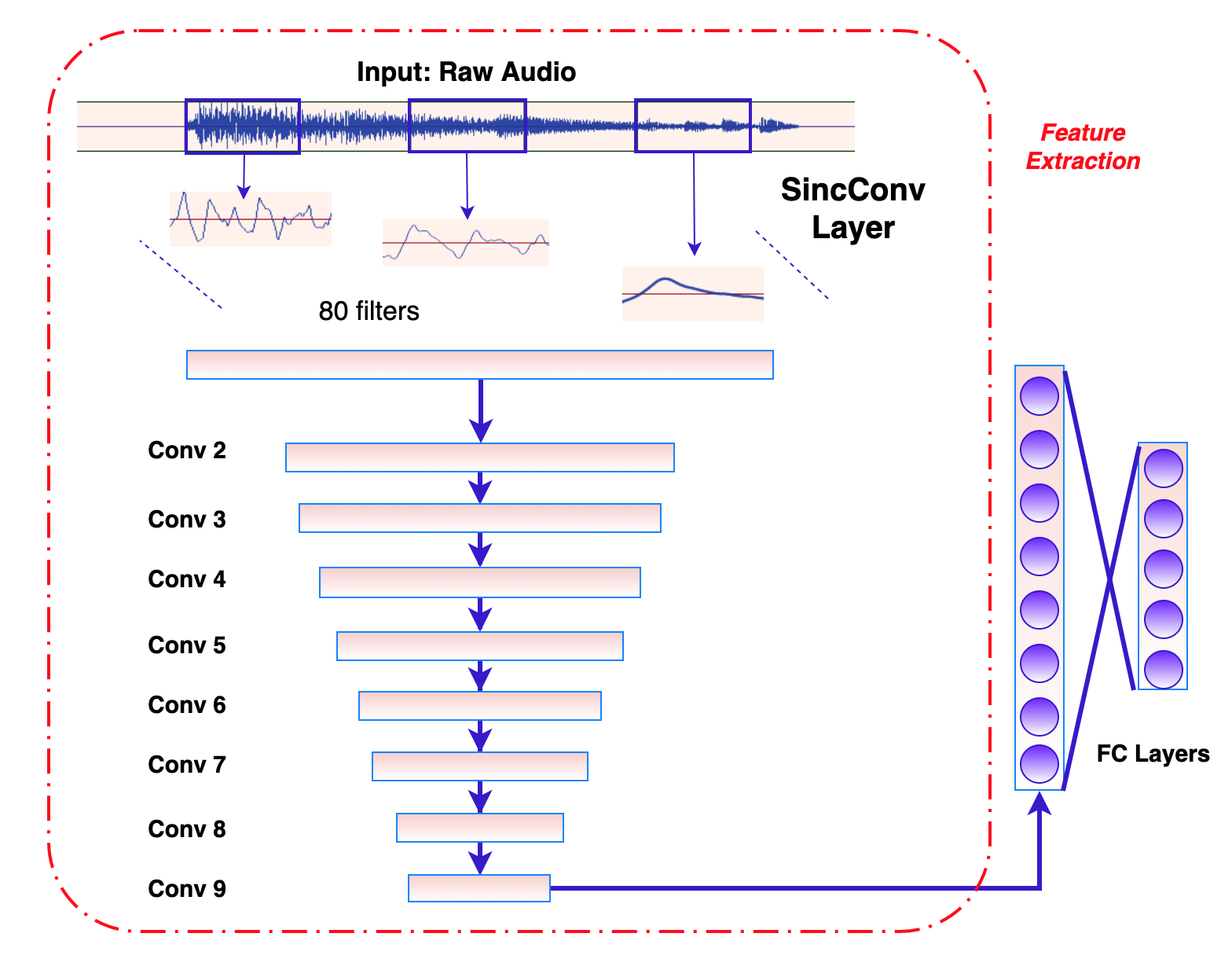


图 3.1基于分类任务下的SCNN网络结构图

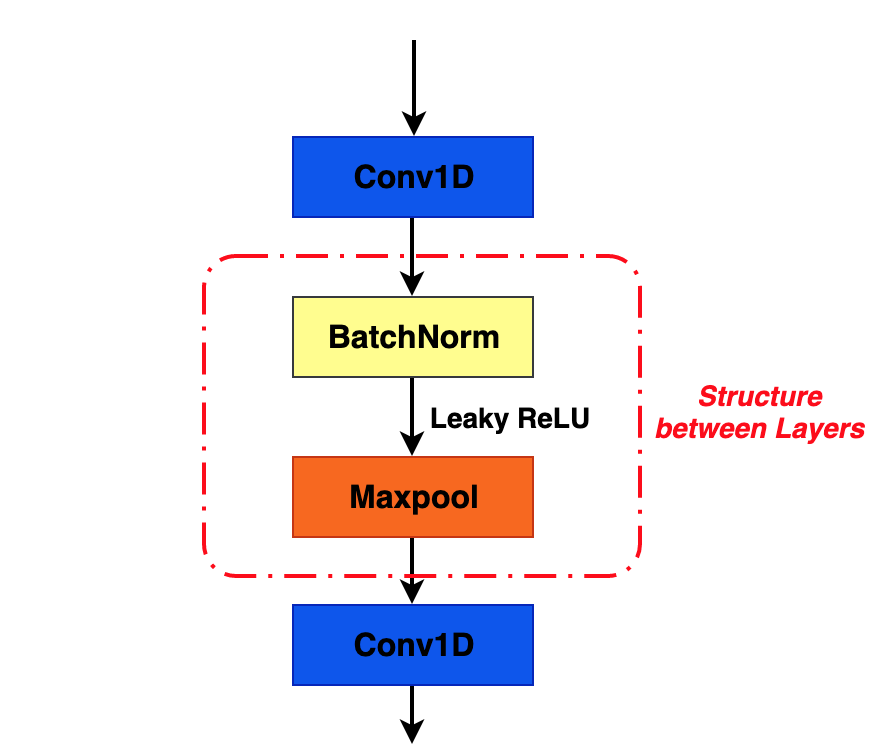


图 3.2 SCNN卷积层间结构图

* 1. SincConv卷积层概述



图 3.3 标准卷积层示意图

图 3.3为标准卷积层操作示意，左侧为输入序列，右侧为卷积核。在该过程中，标准卷积层相当于将原始波形信号与一系列有限长单位冲激响应滤波器(FIR)[33]进行时序卷积，可以用式表示卷积计算过程。

式中的代表一串音频信号，代表长度为的滤波器，然后代表滤波器输出。在标准的卷积层结构中，每个滤波器的所有参数都是从数据中学习到的，但是这些滤波器的可解释性就显得很匮乏，无法很好的表示学习到的个参数的具体意义。因此，SincConv卷积层提出一种由更少参数表示的卷积核，不再学习个参数，如式所示。

式中的代表预定义的函数，仅仅取决于几个可学习的参数。进而，为了更加适合解决音频问题，SincConv卷积层借鉴了数字信号处理中的标准滤波器的方式，定义函数来表示由矩形带通滤波器组成的滤波器组。在频域中，带通滤波器可以表示为两个低通滤波器的差值，如式所示。

式中的和是学习到的低频截止频率和高频截止频率，代表频域上的矩形函数。在回到时域后（使用逆傅立叶变换[33]），该函数变为：

这里的函数定义为：

截止频率可以在范围内随机初始化，其中代表输入信号的采样频率。一种比较好的初始化方式，滤波器可以采用梅尔滤波器组的截止频率初始化，这种设计更加考虑到人的听觉特点，在低频区域设置了更多滤波器。这里的不需要限制在奈奎斯特频率以下，因为在训练的过程中和作为参数可以自动满足该条件。而且，该层网络不需要学习每个滤波器的增益，这些参数可以在后续网络层学习得到。

理想的带通滤波器（即通带完全平坦并且阻带衰减无限的滤波器）需要参数的个数无限多。因此，任何函数都只能实现理想带通滤波器的近似，因为其在通带处存在波动，同时阻带的衰减有限。为了减轻这种频谱能量泄漏的问题，通常会采用加窗的方法[33]，通过将截断函数与窗口函数相乘来实现窗口化，窗口函数可以平滑边缘的突然不连续性，如式所示。

SincConv卷积层中使用的是Hamming窗[34]，定义如下：

Hamming窗特别适合实现高频选择性[34]。但是，实验结果[29]显示SincConv层在采用其他窗函数时没有显着的性能差异，例如Hann，Blackman和Kaiser窗口。SincConv层中涉及的所有操作都是完全可区分的，并且滤波器的截止频率可以使用随机梯度下降（SGD）或其他基于梯度的优化算法与其它层CNN参数一起优化。

第一层SincConv卷积层可以看作将原始音频信号通过若干个带通滤波器，输出的是不同频段的时域信息，以供后端CNN网络进一步进行特征提取。图 3.6为几个学习到的滤波器的时域和频域响应图示例。图 3.4为训练的SCNN模型学习到的所有滤波器的频域响应图，可以直观地看到学习到的不同滤波器的频段响应。

同时，将第一层SincConv卷积层替换为标准卷积层，采用相同的卷积层设置(即80个滤波器，相同的卷积核大小，相同的步长stride以及相同的填充padding)，后端与SCNN结构相同。训练这样的CNN模型，图 3.7为此CNN模型学习到的3个滤波器的时域和频域响应图示例。

通过将本文提出的SCNN模型和标准CNN模型的对比可以发现，由于第一层SincConv卷积层的引入，网络的可解释性显著增强：SCNN第一层起到了带通滤波器的功能，频域响应图可以清晰地显示其频带范围，后续CNN网络相当于对不同频段信息进行降维和整合，提取更深层次的声学特征；而标准CNN模型首层无法学习到明显的可解释性信息，相对表现较差。

为了更明显地反映SincConv层的滤波效果，在一段音频数据上(图 3.5绘制了其时频域图像)，通过上述训练好的滤波器组可以得到不同的音频信息。图 3.8绘制了通过其中几组滤波器后的时域波形和对数梅尔频谱图，来观察其特性。可见，通过第1个滤波器后，音频信号被过滤成低频信号(50Hz左右)；通过第41个滤波器后，音频信号被过滤成中频信号(3000Hz左右)；而通过第71个滤波器后，音频信号被过滤成高频信号(8000Hz左右)。每个滤波器具有频率选择性，原始音频通过SincConv层后相当于将其按照不同频段范围分开，与普通卷积层相比可解释性更好，利于后续网络处理。这里作为比较，将该音频数据通过训练好的CNN模型的几组滤波器，图 3.9显示了其时域波形和对数梅尔频谱图。可以发现，CNN模型的不同的滤波器在各个频段均有分布，缺乏解释性。

* 1. 后端CNN网络设计概述

网络结构如图 3.1，第2层至第9层CNN网络借鉴VGGNet[24]的设计，采用卷积核大小为3的卷积层[46]，随着层数的增加，滤波器数目不断增加(如64,128,256,512)。同时，层与层之间采用Batch Normalization进行批标准化，非线性部分采用Leaky ReLU的激活函数进行计算，目的是为了解决梯度消失和梯度爆炸的问题，使得第一层参数更易于学习与更新。每层间采用不同大小的池化层，进行降维，同时增加网络的泛化能力。3.4节将详细介绍相关方法。

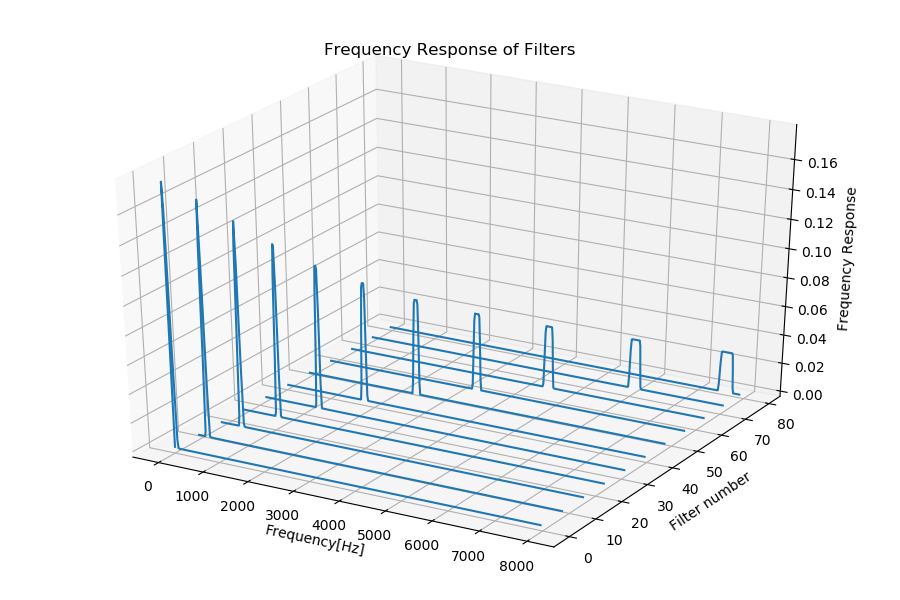


图 3.4 数据集ESC-50下SCNN模型学习到所有80个滤波器的频域响应图，x轴为频率(Hz)，y轴为滤波器数目(每8个显示)，z轴为频率响应。

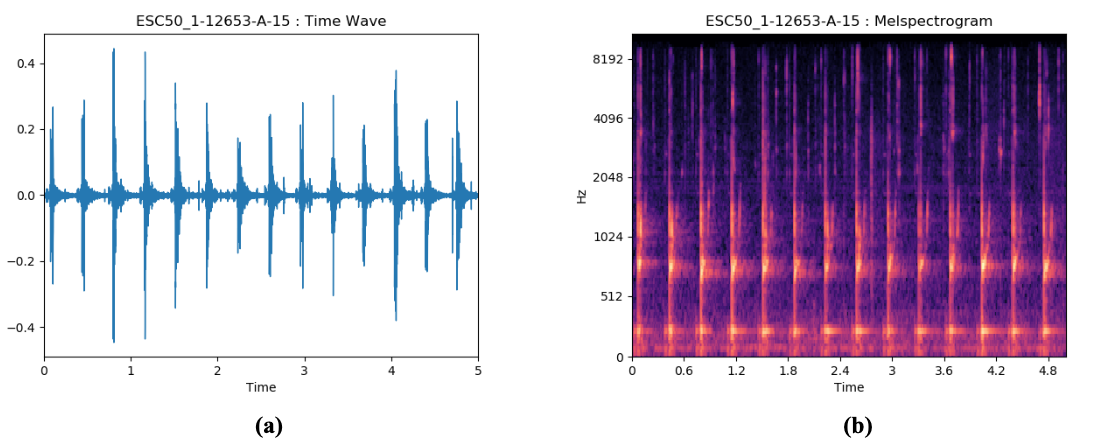


图 3.5 数据集ESC-50上一组数据的时域波形(a)和对数梅尔频谱图(b)

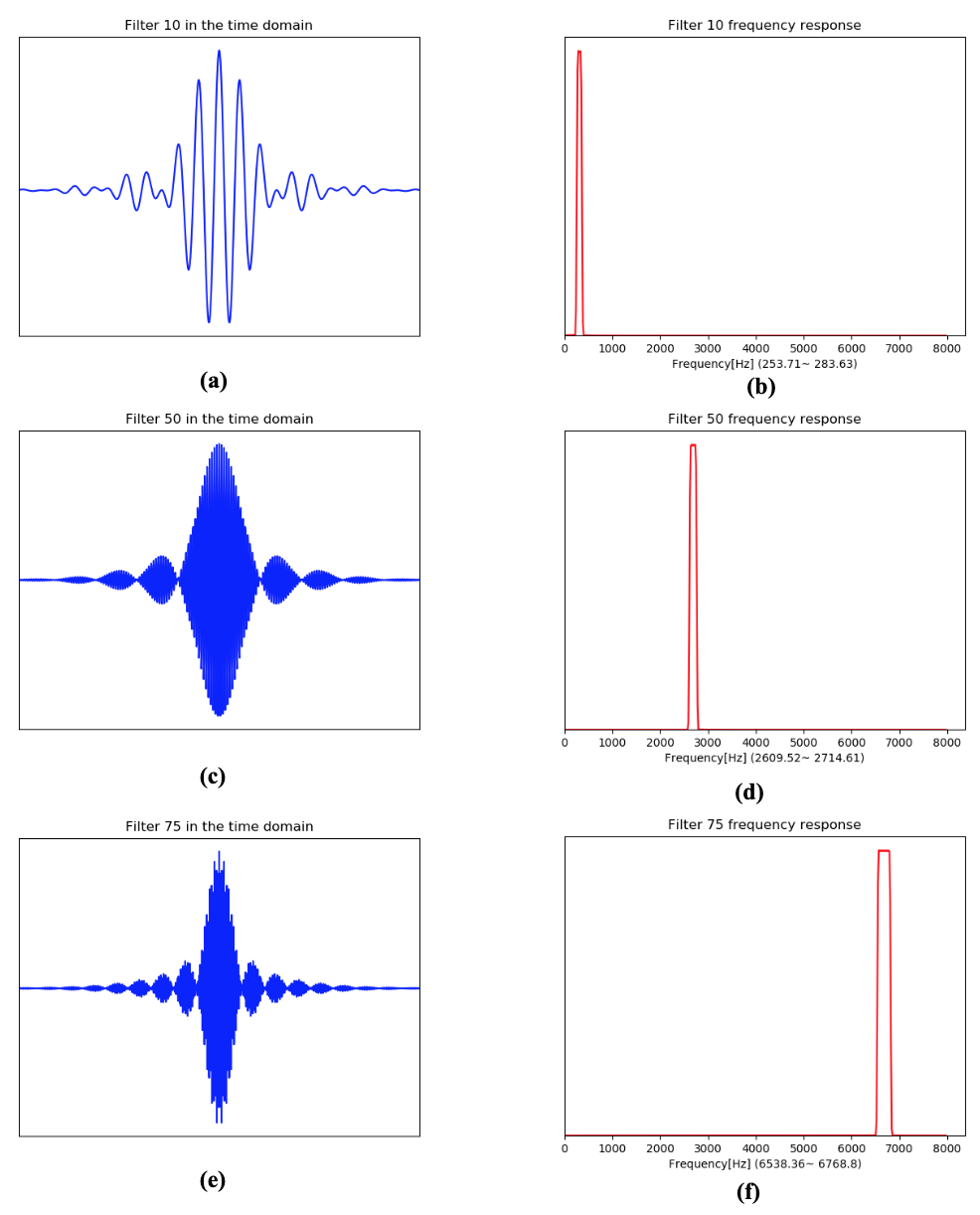


图 3.6 数据集ESC-50下SCNN模型学习到的3组滤波器的时域和频域响应图。a图为第10个滤波器的时域图，b图为第10个滤波器的频域图(可见频带范围为253.71Hz～283.63Hz)，c图为第50个滤波器的时域图，d图为第50个滤波器的频域图(可见频带范围为2609.52～2714.61Hz)，e图为第75个滤波器的时域图，f图为第75个滤波器的频域图(可见频带范围为6538.36Hz～6768.8Hz)。

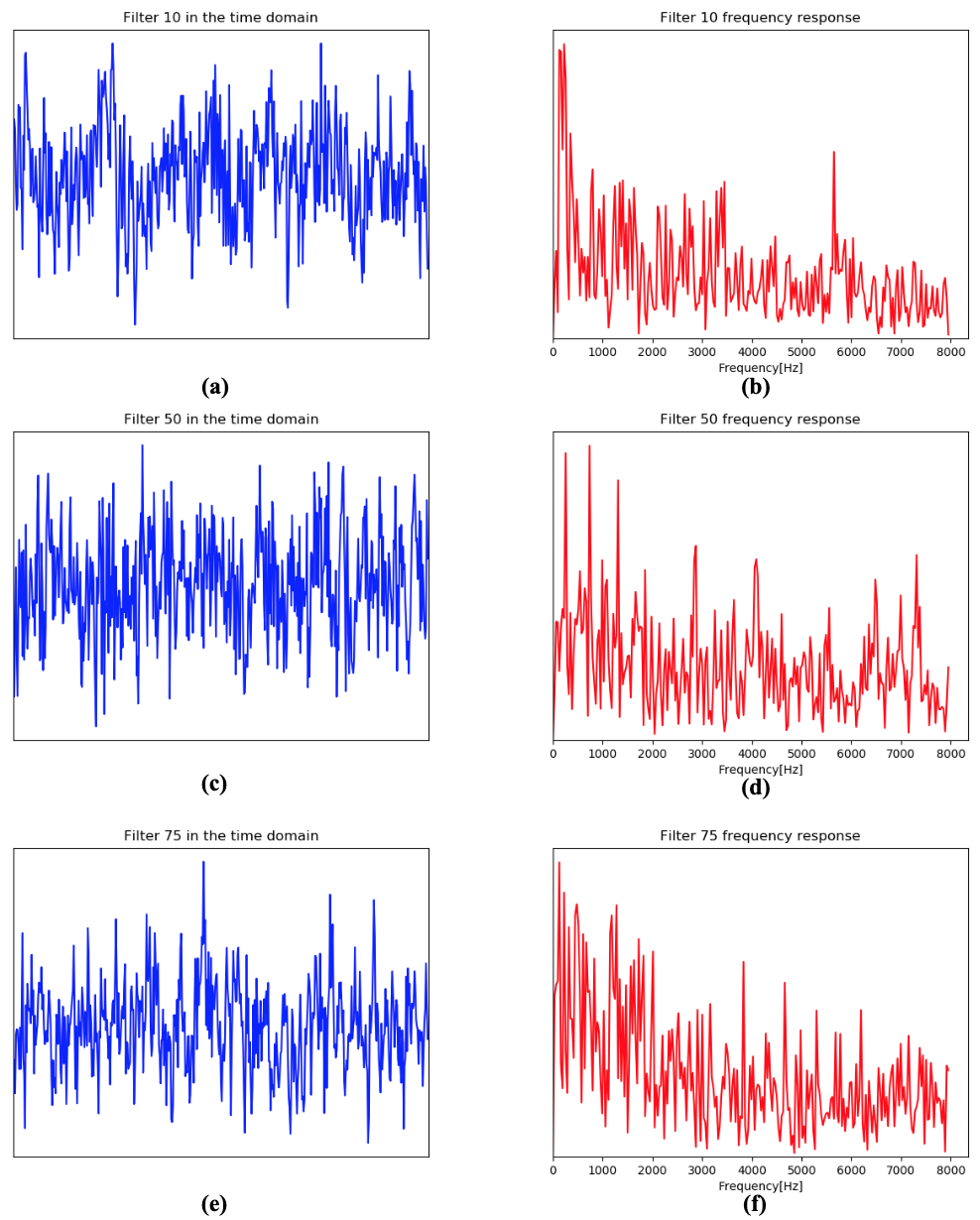


图 3.7 数据集ESC-50下CNN模型学习到的3组滤波器的时域和频域响应图。a图为第10个滤波器的时域图，b图为第10个滤波器的频域图，c图为第50个滤波器的时域图，d图为第50个滤波器的频域图，e图为第75个滤波器的时域图，f图为第75个滤波器的频域图。

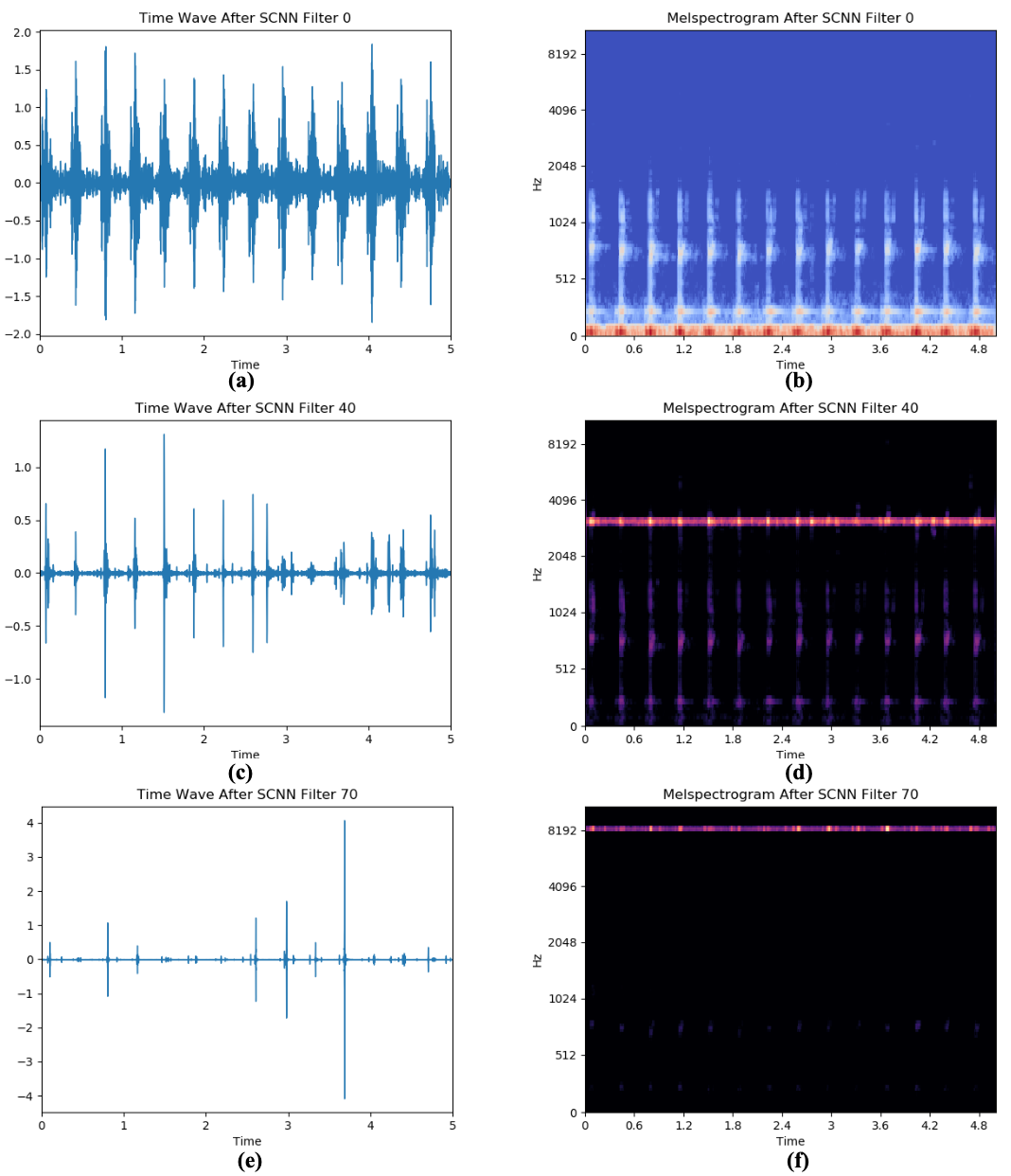


图 3.8 数据集ESC-50上一组数据(见图 3.5)分别通过训练好的SCNN网络模型的第1个、第41个和第71个滤波器后得到的时域波形(a,c,e)和对数梅尔频谱图(b,d,f)

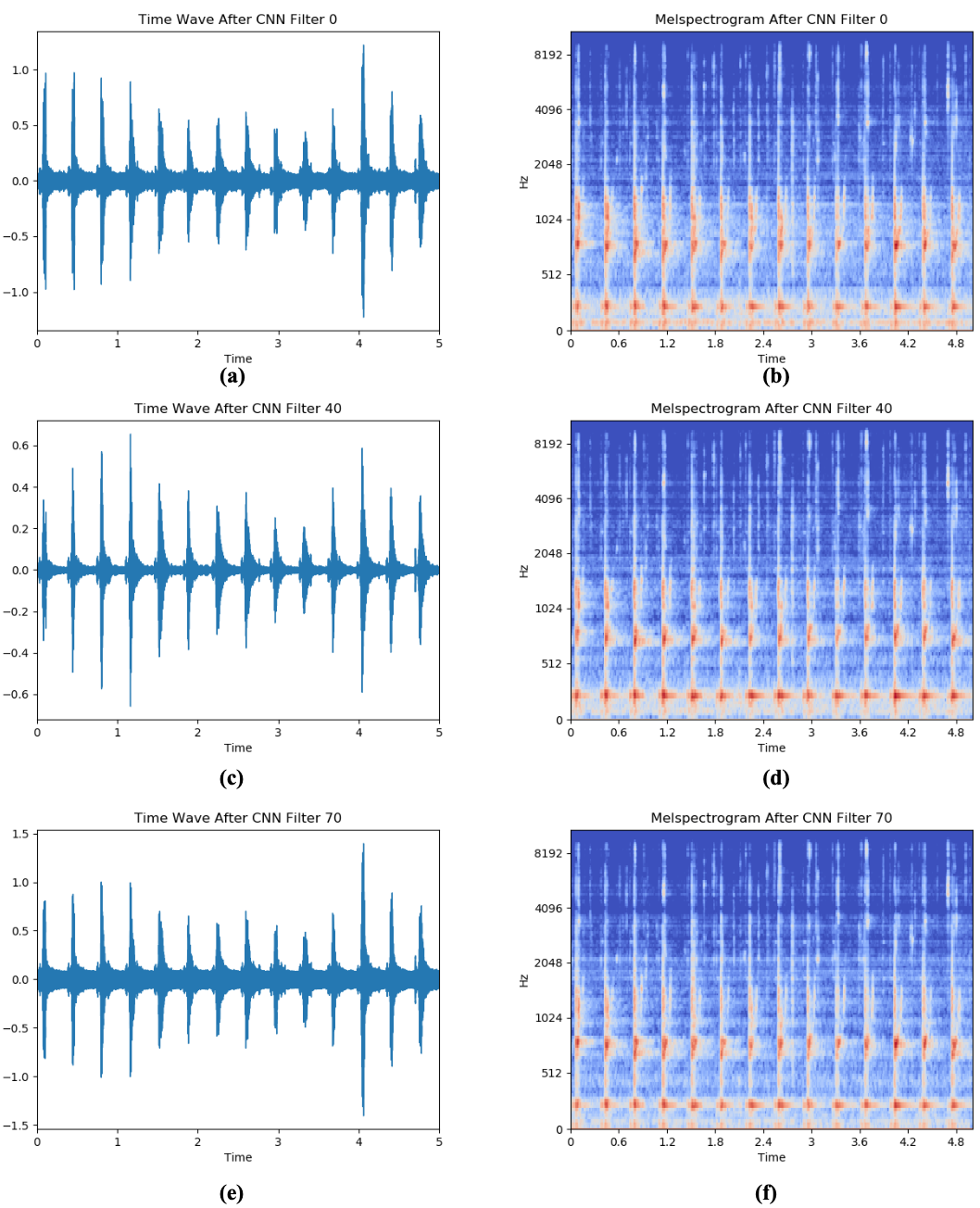


图 3.9 数据集ESC-50[35]上一组数据(见图 3.5)分别通过训练好的CNN网络模型的第1个、第41个和第71个滤波器后得到的时域波形(a,c,e)和对数梅尔频谱图(b,d,f)

* 1. 相关方法介绍
     1. 激活函数

激活函数是反应多层神经网络中上层节点的输出和下层节点的输入之间关系的函数。目前最常用的激活函数是ReLU函数，其与sigmoid函数和tanh函数相比，在一定程度上解决了梯度消失问题，计算速度和收敛速度都更快。然而，由于ReLU对于负值的处理方式是将其置为0，该神经元之后的梯度就永远是0了，即出现Dead ReLU问题。这种问题对于选择了不合理的学习率的情况尤为严重。

解决Dead ReLU问题的一种有效方式是使用Leaky ReLU激活函数，如所示。

通过将ReLU的前半段设为而不是0，通常设置。本文采用Leaky ReLU激活函数，并且网络参数采用Kaiming初始化的方式。

* + 1. 损失函数

SCNN网络采用交叉熵损失函数，其计算公式如式所示。

式中，是类别的真实值，采用one-hot编码形式；是类别的预测值，此处是经过softmax后的结果。由于非0即1，在Pytorch代码中，目标不是以one-hot编码形式表示，而是scalar表示，可以直接使用F.cross\_entropy计算交叉熵，实质上调用了log\_softmax和nll\_loss，其计算方式如式所示。

* + 1. 优化算法

由于在训练的过程中把数据分批进行训练，常用的算法是mini-batch SGD。但是这种算法可能出现在最优点附近徘徊的情况。同时需要设置一个合适的学习率，当学习率较小时，网络在训练的时候收敛太慢；当学习率较大时，会导致在训练过程中可能跳过最优点。一种基于梯度的移动指数加权平均的方法可以在一定程度上缓解上述问题，称为Momentum优化算法。本文采用Momentum优化算法：

其中，和分别是损失函数在前轮迭代过程中累积的梯度动量，是梯度累积的指数，通常设置值为0.9。 和分别是损失函数反向传播时所求得的梯度，是网络的学习率。使用Momentum优化算法可以解决mini-batch SGD优化算法更新幅度摆动大的问题，同时可以使得网络的收敛速度更快。

* + 1. 梯度问题和过拟合问题的处理

梯度消失和梯度爆炸问题是深度神经网络训练时必须解决的问题。本文除了采用了Leaky ReLU激活函数外，还在层与层之间使用了Batch Normalization批标准化(简称BN)以解决梯度问题。同时，BN还可以提升训练速度，加快收敛，并对于初始条件要求不那么高，即允许使用较大的学习率。

BN还能增加分类效果，可理解为一种类似于Dropout的防止过拟合的正则化表达方式。本文在SCNN网络后的全连接层里(如图 3.1)使用Dropout方法防止过拟合。

* 1. 本章小结

本章介绍了一种基于CNN的深度声学特征网络结构，即SCNN网络。具体叙述了其设计思路和设计细节。SCNN网络是下文声学事件检测和声学场景分类系统以及说话人系统的前端特征提取网络，第四章和第五章将详细介绍这些系统。

1. 基于SCNN网络的声学事件检测和声学场景分类系统
   1. 引言

本章介绍了基于SCNN网络的声学事件检测和声学场景分类系统，4.2节概述了声学事件检测任务和声学场景分类任务，4.3节和4.4节详细介绍系统设计情况，4.5节介绍了相关实验及结果分析。

* 1. 声学事件检测和声学场景分类任务概述

从广义上讲，声学事件[12]一词是指由不同的物理声源产生的特定声音，例如经过的汽车声，鸟鸣或门铃等。声学事件通常只有一个来源，尽管如汽车的声音同时来自于车轮和引擎等，定义用什么来作为单一的来源仍然是一种主观选择。声学事件通常具有明确的定义和短暂的持续时间。相比之下，声学场景[12]是指当来自各种源的声音（通常来自真实场景）混合时形成的声音的整体。例如，街道的声学场景可以包含经过的汽车，脚步声，人们说话等。家中的声学场景可能包含来自收音机的音乐，洗碗机嗡嗡声和儿童大喊大叫等。

声学场景和事件的计算分析的总体目标是通过计算方法从音频中提取信息，所要提取的信息类型取决于具体的应用。但是，我们可以将典型的声音分析任务分类为几个高级别的类别。然后分类任务的目标是将音频记录分类为一组（预定义）类别中的一个。例如，声学场景分类系统可以将音频分类为包括家庭，街道和办公室的一组类别之一。而在声学事件检测中，则可以在声音发生时，找到一种或多种声学事件，并确定其出现的时间位置。还有其他更具体的任务，例如估计两个录音是否来自同一个声场。

声学场景和事件的分类和检测[37][38][39]对于多媒体检索，基于音频的监视和监测等具有重要作用。由于每种声音的时间和频率特性变化很大，以及音频中可能的附加背景噪声、声学事件之间的重叠等原因，声学事件检测和声学场景分析存在着很大的挑战。解决这两个问题的系统的整体性能依赖于两个部分：特征提取部分和分类部分。一个好的特征应该能够有效区分不同类别的声音，并且在给定的音频类别内变化很小，同时，应该具有较好的抗噪能力。

* 1. 基于SCNN网络的声学事件检测系统概述

本文提出一种基于SCNN网络的声学事件检测系统，如图 4.1。特别的是，音频事件检测任务通常不会将全部音频作为输入，通常取1s～2s的音频[46]。原因是选取时长过长，往往会有很长时间的静默或无用信息，造成参数量过大，浪费内存；选取时长过短，可能判断依据不足，例如无法分辨敲击键盘声和点击鼠标声。本文以输入音频信号为例(采样频率为16kHz，时长1.5s)，通过SCNN特征提取网络输出的深度特征维数为，之后是两层全连接层用于分类。

在第一层SincConv卷积层[29]，采用的滤波器数目为80，卷积核大小为81个单位长度(约5ms)，卷积步长为1个单位长度。这里借鉴了梅尔滤波器的设计思路，即对于4kHz带宽的语音信号一般采用24个滤波器，而对于8kHz带宽的语音信号一般会采用40个滤波器[43]。滤波器的数目越多，即带通滤波器频段范围越小，所得到的时域信息解释性越强，但是每个频段之间的信息关联性变差，例如在0～8kHz范围内取8k个滤波器，那么相当于每个Mel标度都有一个带通滤波器进行学习，但是很难反应出这些滤波器学习到的信息之间的关系，而且计算过程浪费了大量内存；反之，滤波器的数目越少，每个带通滤波器的频段范围越大，频段内部的信息关联性越强，然而所得到的时域信息解释性变差，例如考虑一个极端，当滤波器个数为1时，就相当于一个普通的CNN卷积层，解释性差，带通滤波器的思想就没有起到作用。考虑到上述因素并结合实验分析，最终选择滤波器个数为80。

对于语音信号，可以认为其在15ms～25ms范围内为稳态信号，一般窗口会设置为20ms左右，同时采用10ms的Overlap。但是对于声学事件，信号的稳态时间一般没有那么长，通常认为在5ms左右声学事件的信号处于稳态[44][45]。因此，为了让设计的特征提取器能够在多个任务下工作，这里设置了较小的卷积核(81个单位长度)和卷积步长(1个单位长度)。

后端CNN网络均采用卷积核大小为3的卷积层来进一步提取时域信息，小的卷积核使得每层的参数量减少，以此来控制模型大小，减少计算代价[46]。同时，深层网络增加了非线性单元，使得拟合效果更好。

每一层之间采用了Batch normalization批标准化(BN)。BN在一定程度上缓和了梯度爆炸和梯度消失的问题，这在深度网络结构的优化中是一个重要问题。同时BN也加速了训练过程。激活函数采用了Leaky ReLU，非饱和激活函数在一定程度上可以解决梯度消失问题，能够加快收敛速度。同时，Leaky ReLU解决了ReLU的负值问题。

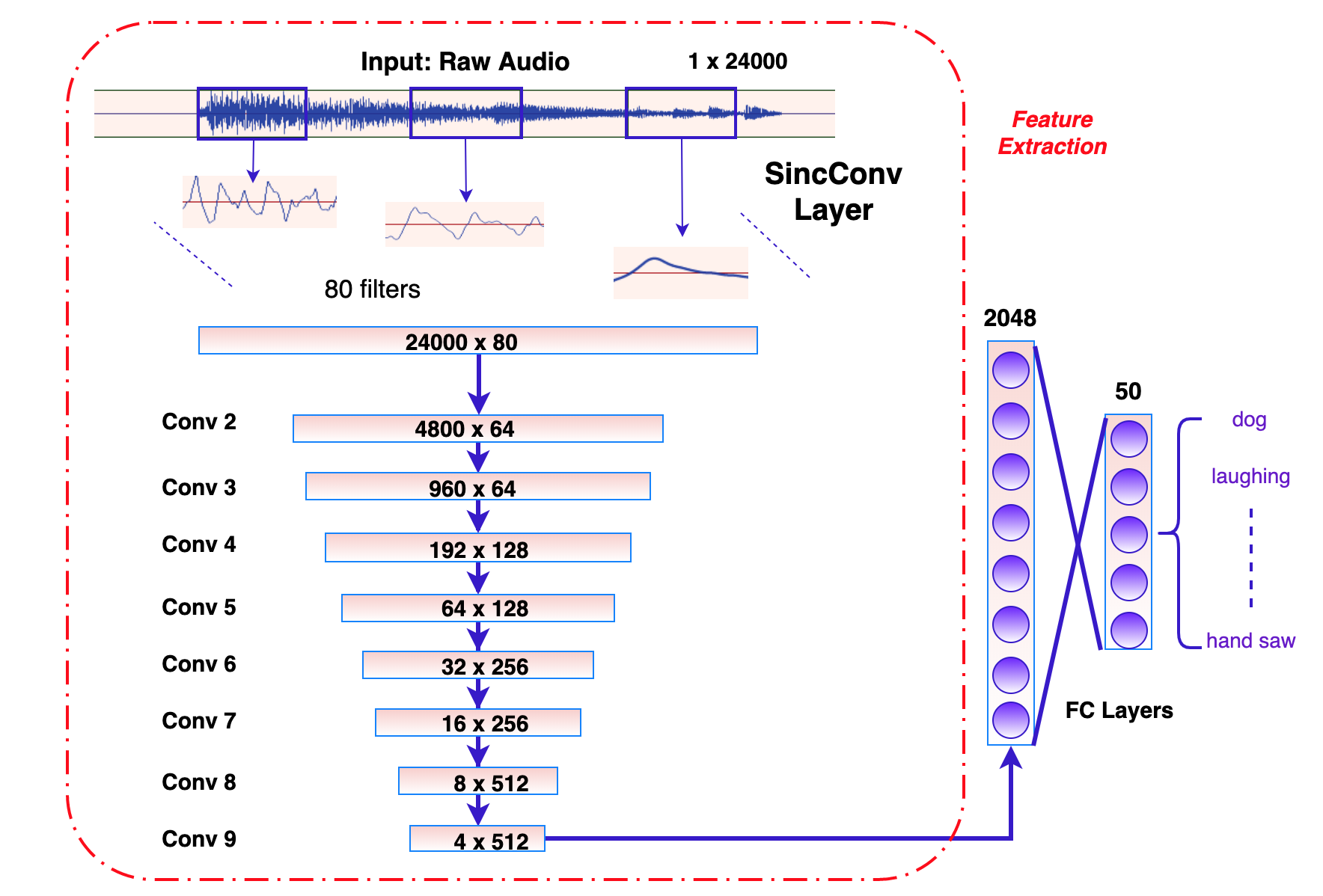


图 4.1 基于SCNN网络的声学事件检测系统示意图(ESC-50数据集[35]下)

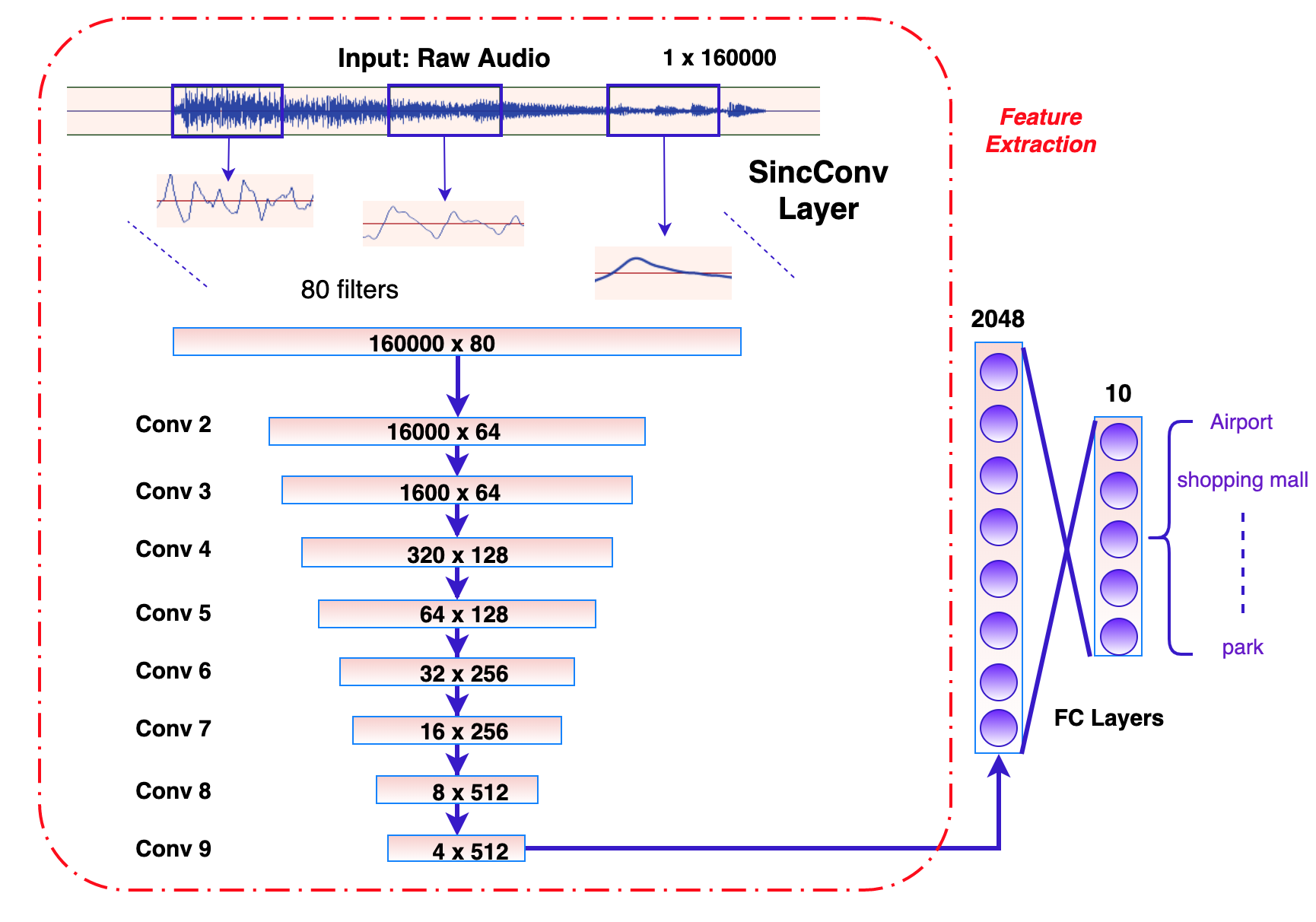


图 4.2 基于SCNN网络的声学场景分类系统示意图(TUT Urban Acoustic Scenes 2018，Development dataset数据集[36]下)

* 1. 基于SCNN网络的声学场景分类系统概述

本文提出的声学场景分类系统与4.3节所介绍的声学事件检测系统基本一致，如图 4.2。唯一不同的是，由于声学场景是长时系统，因此声学场景分类系统的输入使用完整音频，以输入音频信号为例(采样频率为16kHz，时长10s)，通过SCNN特征提取网络输出的深度特征维数为，之后是两层全连接层用于分类。

* 1. 实验设置及结果分析对比
     1. 基于SCNN网络的声学事件检测系统实验设置

1. 数据集

用于声学事件检测系统的训练和测试数据集为ESC-50数据集[35]。

ESC-50是ESC(Dataset for Environmental Sound Classification)中包含50种常见声音事件的2000个短片段的数据集。每类声音事件含有40个数据样本，并由统一格式存储(采样频率为44.1 kHz，单通道，以192kbit/s的速度进行Ogg Vorbis压缩)。为方便介绍，可以将其分为5个主要类别(每个类别10类):

1) 动物的声音；

2) 自然的音景和水声；

3) 人（非语音）声音；

4) 室内/家庭声音；

5) 外部/城市噪音

ESC-50数据集提供了各种类型的声源:一些非常常见（如笑声，猫喵喵叫，狗吠等），一些非常不同（如玻璃破碎，刷牙等），还有一些差异更细微（如直升机和飞机噪音）。该数据集可能存在的缺陷之一是每个类可用的样本数量有限。同时，给出了实验参与者（人类）实际的测试准确率。

2. 系统结构及参数设置

实验采用Pytorch框架，系统具体结构及相关方法已在4.3节介绍。实验过程中，使用5折交叉验证的方法，对于ESC-50数据集，采用随机选取原始音频中1.5s的连续信息作为输入(原始数据每段时长为5s)，在测试阶段，本文采用了probability-voting机制[46]。Batch size设置为64，通过momentum SGD优化器优化网络，momentum设置值为0.9。学习率初始设为0.01，并随着训练的进行采取指数衰减，衰减指数设为0.985。每一层的权重都是从零开始初始化的，没有采用任何预训练模型或者是Gammatone初始化，目的是为了充分测试该模型的性能。最后一层全连接层使用了Dropout层来避免过拟合，Dropout比率设置为0.5。所有的权重参数均服从系数为的L2正则化。Epoch设置为500， 当Epoch为400左右网络基本训练完毕。

* + 1. 基于SCNN网络的声学场景分类系统实验设置

1. 数据集

用于声学事件检测系统的训练和测试集为TUT Urban Acoustic Scenes 2018，Development dataset[36]。

TUT Urban Acoustic Scenes 2018, Development dataset数据集是在DCASE(IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events)2018年竞赛中的Task 1(Acoustic scene classification)的Subtask A使用的数据集。DCASE是基于声学场景和声学事件的检测和识别任务提出的竞赛挑战，共包括5项任务。

Task 1的数据集是TUT Urban Acoustic Scenes 2018数据集，包括来自各种声学场景的记录。该数据集记录在六个欧洲大城市，每个场景类别的不同位置。对于每个录制位置，有5-6分钟的音频。原始录音被分成长度为10秒的片段。有关录制内容的可用信息包括：声学场景类，城市和录制位置。该数据集包含以下10类声学场景：

1. 机场-airport；
2. 室内购物中心-shopping mall；
3. 地铁站-metro station；
4. 步行街-street pedestrian；
5. 公共广场-public square；
6. 街道交通-street traffic；
7. 乘坐电车-tram；
8. 乘坐公共汽车-bus；
9. 乘坐地铁-metro；
10. 城市公园-park。

该数据集使用四个同时捕获音频的设备进行录制。主要录音设备包括Soundman OKM II Klassik /工作室A3，驻极体双声道麦克风和使用48kHz采样率和24位分辨率的Zoom F8录音机。麦克风专门制作成耳机佩戴在耳朵上。作为其效果，所记录的音频非常类似于到达佩戴该设备的人的人类听觉系统的声音。TUT Urban Acoustic Sc​​enes 2018数据集每个声学场景具有864个片段（144分钟的音频）。该数据集总共包含8640个片段，即24小时的音频。

2. 系统结构及参数设置

系统具体结构及相关方法已在4.4节介绍。实验过程中，使用DCASE Task1给定的训练集和测试集，输入为整段音频(每段时长为10s)。Batch size设置为32，通过momentum SGD优化器训练网络，momentum设置值为0.9。学习率初始设为0.01，并随着训练的进行采取指数衰减，衰减指数设为0.985。每一层的权重都是从零开始初始化。最后一层全连接层使用了Dropout层来避免过拟合，Dropout比率设置为0.5。所有的权重参数均服从系数为的L2正则化。Epoch设置为800，当Epoch为700左右网络基本训练完毕。

* + 1. 实验结果分析及对比

1. 实验结果分析

表 4.1展示了声学事件检测系统下SCNN网络性能的测试结果。同时，将其一些其它的CNN模型进行对比。可以发现，所提出的SCNN网络性能明显优于传统的Logmel方法(66.50%)和其它的CNN方法，并且接近于人类区分率(81.30%)。

|  |  |
| --- | --- |
| 表 4.1 Comparison of Accuracy of ESC-50 Dataset | |
| Model | Accuracy(%) |
| Piczaks CNN | 64.50 |
| Tokozumes Logmel-CNN | 66.50 |
| EnvNet | 64.00 |
| AlexNet | 69.00 |
| GoogLeNet | 73.20 |
| WaveMsNet | 70.05 |
| **SCNN (ours)** | **78.10** |
| Human performance | 81.30 |

表 4.2展示了声学场景分类系统下SCNN网络性能的测试结果，并与Baseline和WaveMsNet进行对比，效果优于二者。

图 4.3和图 4.4展示了ESC-50数据集训练过程中的Loss和Accuracy变化情况，可以看到Epoch至400左右时，网络基本收敛，训练完成。图 4.5和图 4.6分别为两个数据集下分类任务的混淆矩阵。

|  |  |
| --- | --- |
| 表 4.2 Comparison of Accuracy of DCASE2018 Task1 | |
| Model | Accuracy(%) |
| Baseline | 61.00 |
| WaveMsNet | 69.50 |
| **SCNN (ours)** | **69.80** |

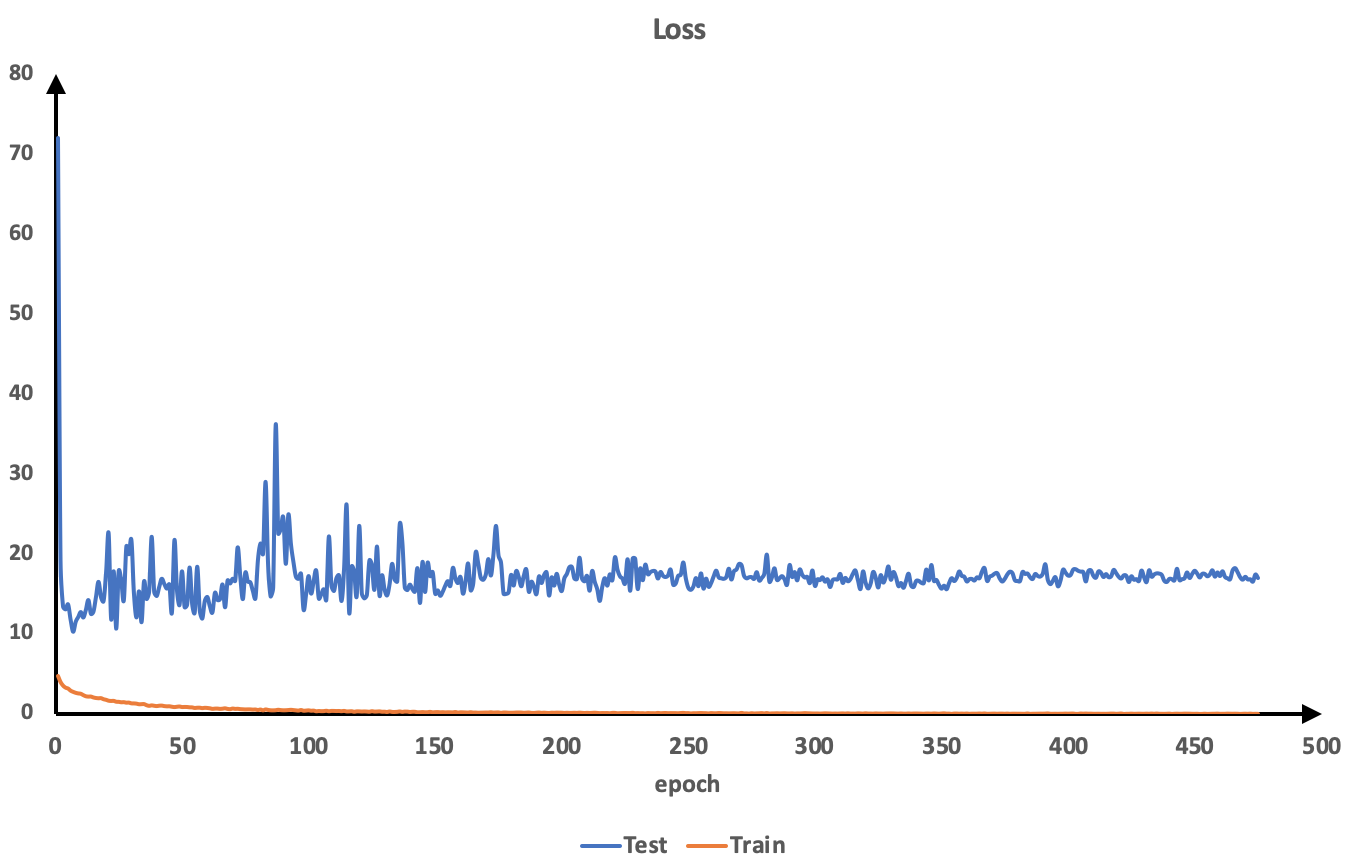


图 4.3 训练阶段Loss情况

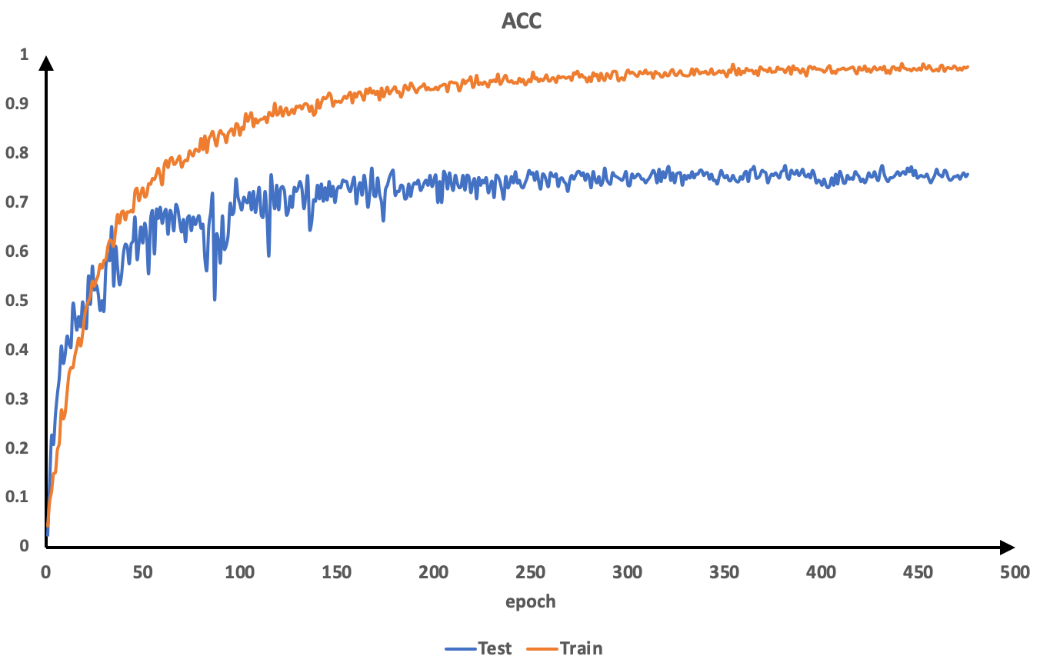


图 4.4 训练阶段Accuracy情况

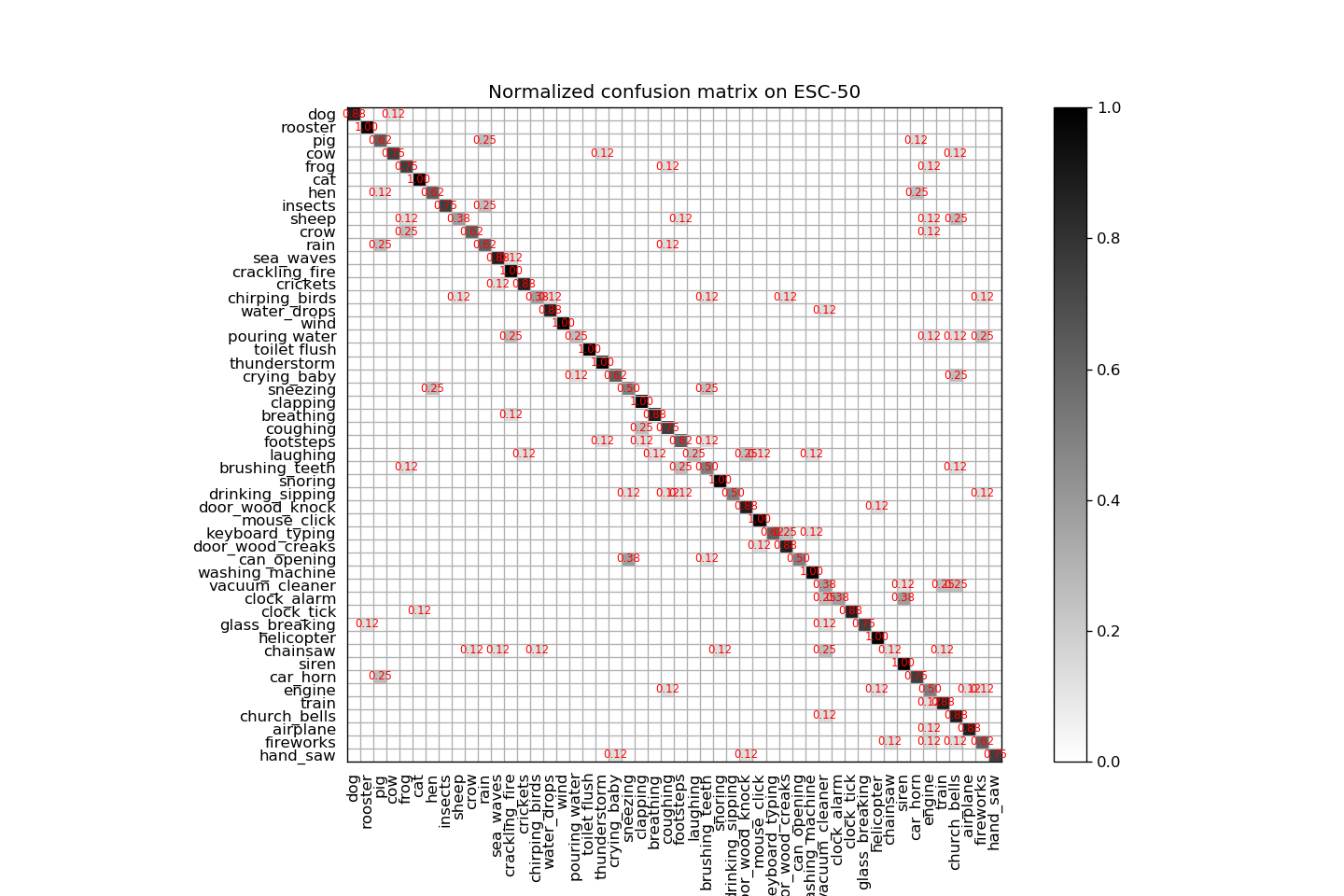


图 4.5 ESC-50数据集上混淆矩阵

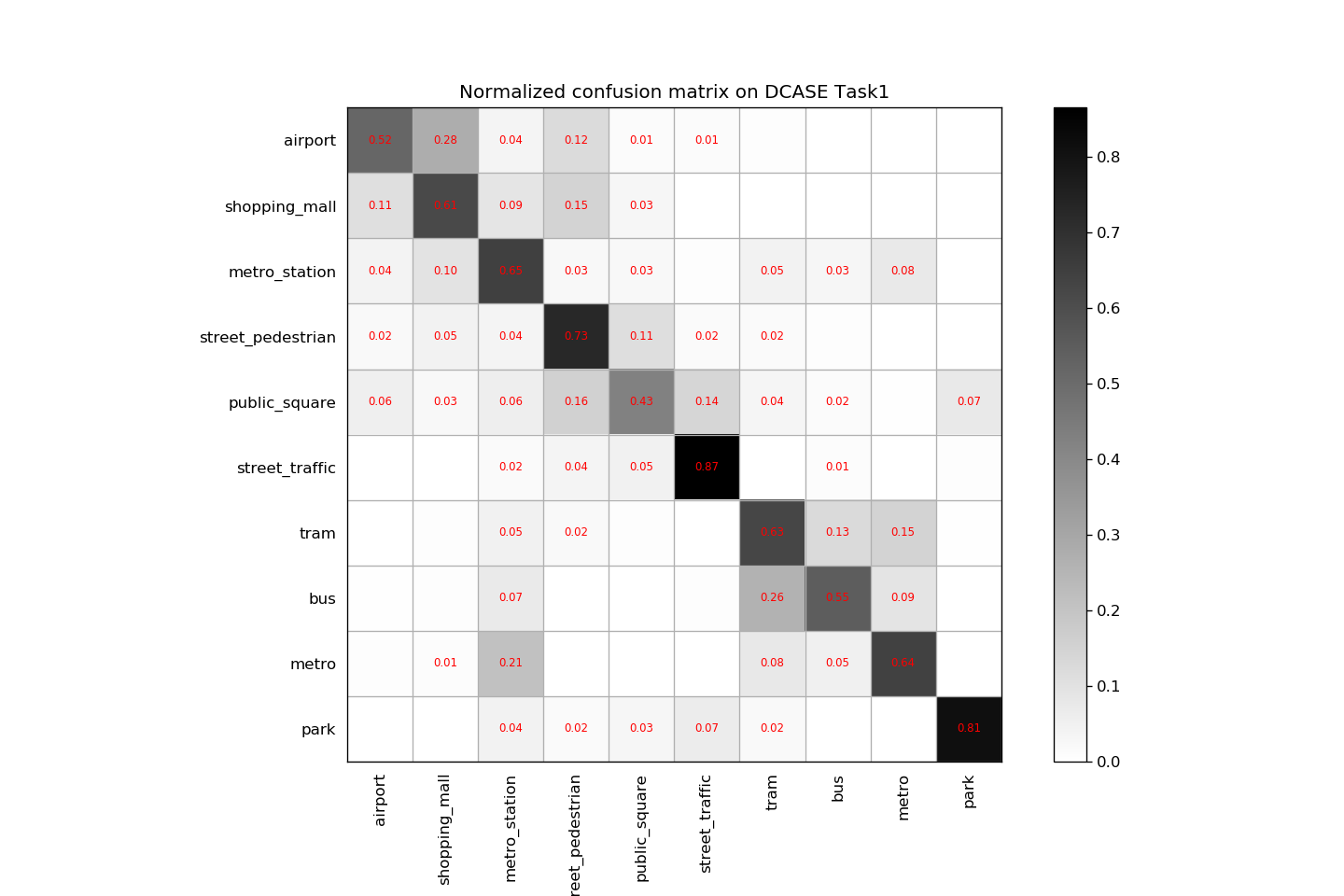


图 4.6 DCASE Task1数据集上混淆矩阵

2. 系统对比分析

为了进一步体现该声学特征模型(SCNN)的特性，我绘制了其在ESC-50数据集和DCASE2018 Task1数据集(TUT Urban Acoustic Scenes 2018, Development dataset)下的累积频域响应图，如图 4.7。累积频域响应图[44]是通过对学习到的每一个滤波器的频域响应进行叠加得到的，可以比较直观地看到不同数据集的音频频段的大致范围。从图 4.7可以看出，SCNN网络在两个数据集上学习到的低频段比重更大，这符合人耳对于1000Hz左右的低频信息更敏感的特性。不同的是，在ESC-50数据集上学习到的频域范围在8000Hz以内几乎都有分布，而在DCASE2018 Task1数据集上学习到的频域范围集中在5000Hz以内。我们可以通过绘制训练过程的频域累积变化图来更加直观地观察这一现象，如图 4.8。

为了分析这一现象的原因，在两个数据集上随机选取两组数据绘制其对数梅尔频谱图，如图 4.9。可以发现，ESC-50数据集的数据在8000Hz以内都有频谱分布，而DCASE2018 Task1数据集的数据主要集中在4000Hz以内。该实验进一步反映了SCNN网络有效学习到了不同数据集的频段特性。

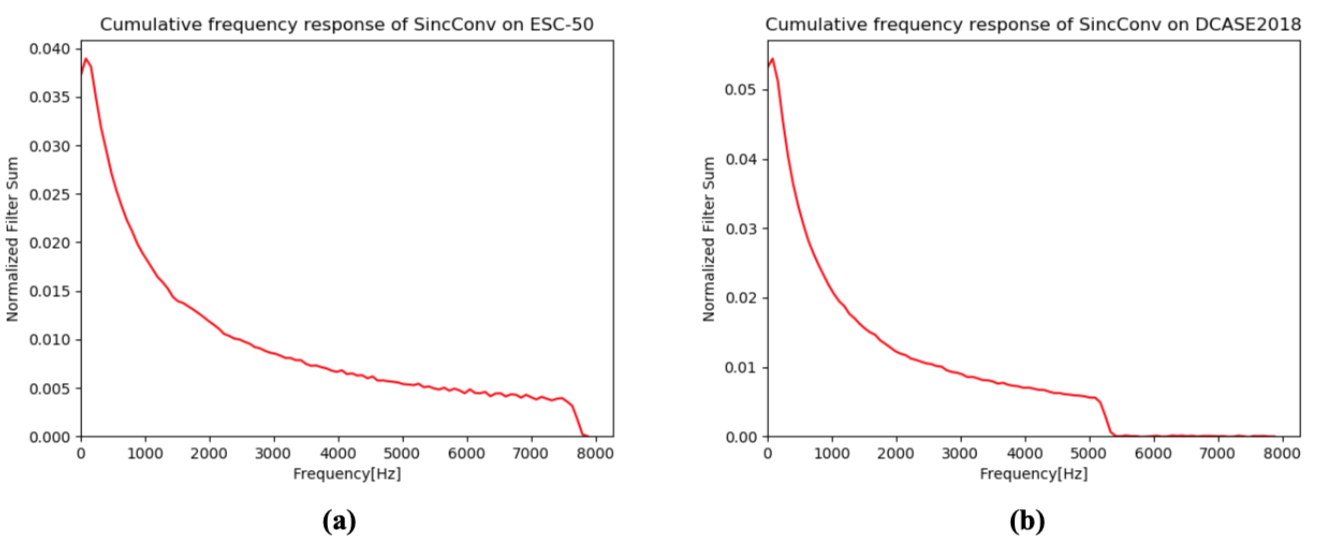


图 4.7 SincConv卷积层在ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)下学习到的累积频域响应图。

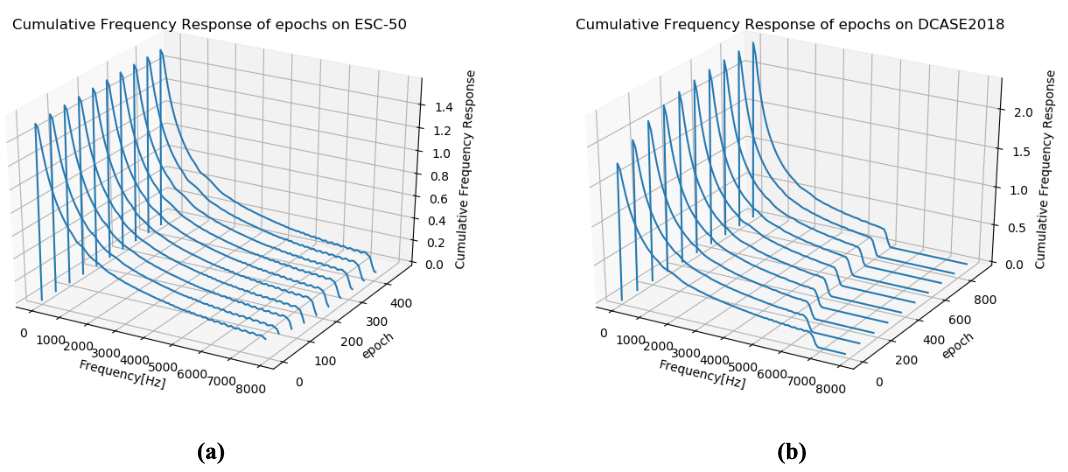


图 4.8 随着训练的进行，在ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)下学习到累积频域响应图变化过程。x轴为频率(Hz)，y轴为Epoch(每50个显示)，z轴为频率响应。



图 4.9 ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)中随机两组数据的对数梅尔频谱图

* 1. 本章小结

本章介绍了基于SCNN网络的声学事件检测和声学场景分类系统，并设置了实验对其性能进行评估。实验结果表明，SCNN网络在声学事件检测任务和声学场景分类任务上性能优于一些主流CNN模型，所提出的端到端方法效果显著。同时，SCNN网络很好地适应不同的数据集和分类任务，能够有效学习声学特征。

1. 基于SCNN网络的说话人识别系统
   1. 引言

本章介绍了基于SCNN网络的说话人识别系统，5.2节概述了说话人任务，5.3节详细介绍系统设计及系统结构，5.4节介绍了相关实验及结果分析。

* 1. 说话人识别任务概述

通过他或她的声音来识别一个人，是人与人之间的自然交流的重要保证。例如在电话交谈中，通常需要首先确定谁在说话，或者至少是在熟悉的发言者的情况下，听者认为发言者身份正确，谈话才能正常进行。自动说话人识别[40][41]系统已经成为许多电子商务应用程序，商业交互，取证和执法中验证身份的重要手段。

受过法医说话人识别训练的人类专家可以通过在称为结构化听力的普适方法中检测语音中的一组声学，韵律和语言特征来更好地完成这项任务。法医语言科学家和语言学家多年来开发了法医说话人识别技术，来帮助减少任何潜在的偏见或对潜在嫌疑人的未知音频样本和参考模板的有效性的先入为主的理解。经验丰富的信号处理和机器学习研究人员继续开发了自动算法，以有效地执行说话人识别任务。主要分为两类任务：说话人识别和说话人验证。

在说话者识别中，任务是从一组已知说话者中识别未知说话者。换句话说，目标是找到来自音频样本内的听起来最接近于未知说话者的语音的说话者。

在说话者验证中，未知说话者声称拥有身份，而任务是验证此声明是否属实。这基本上归结为比较两个语音样本/话语并确定它们是否由相同的发言者说出。

每个说话人在他或她的声音中都有一些独特的特征，主要是由于说话人的声道生理学和发音的清晰程度不同。个别说话者的特征可能不容易区分，比如同卵双胞胎具有相似的声道形状和声学特性，并且难以将它们与感知或取证视角区分开来，但是其声音还是存在着不同之处。因此，无论识别是由人（专家或普通人）还是由机器执行，都需要考虑一些可测量和预定义的语音方面的信息，以便在语音之间进行有意义的比较。通常，我们将这些表征方面称为特征参数[41]。

人们可能会认为独特的声音必须具有独特的功能，但这并非总是如此。例如，两个不同的说话者可能具有相同的说话率（这是一个有效的特征参数），但平均音高不同。这就是为什么考虑多个特征参数是至关重要的。说话者识别问题属于典型的模式识别问题，所采用的方法一般由特征提取器、分类器组成。特征提取器的性能很大程度上决定了说话人识别任务的可靠性，因而这一部分至关重要。

* 1. 基于SCNN网络的说话人识别系统概述

本文提出一种基于SCNN网络的说话人系统，如图 5.1。本文以输入音频信号为例(采样频率为16kHz，时长0.4s)，通过SCNN特征提取网络输出的深度特征维数为。说话人识别任务不同于声学事件和声学场景分类，其频域特征比较集中(前者声源是人类，而后者是不同类别)[29]，这里采用四层全连接层构成的DNN网络作为后端网络进行分类。

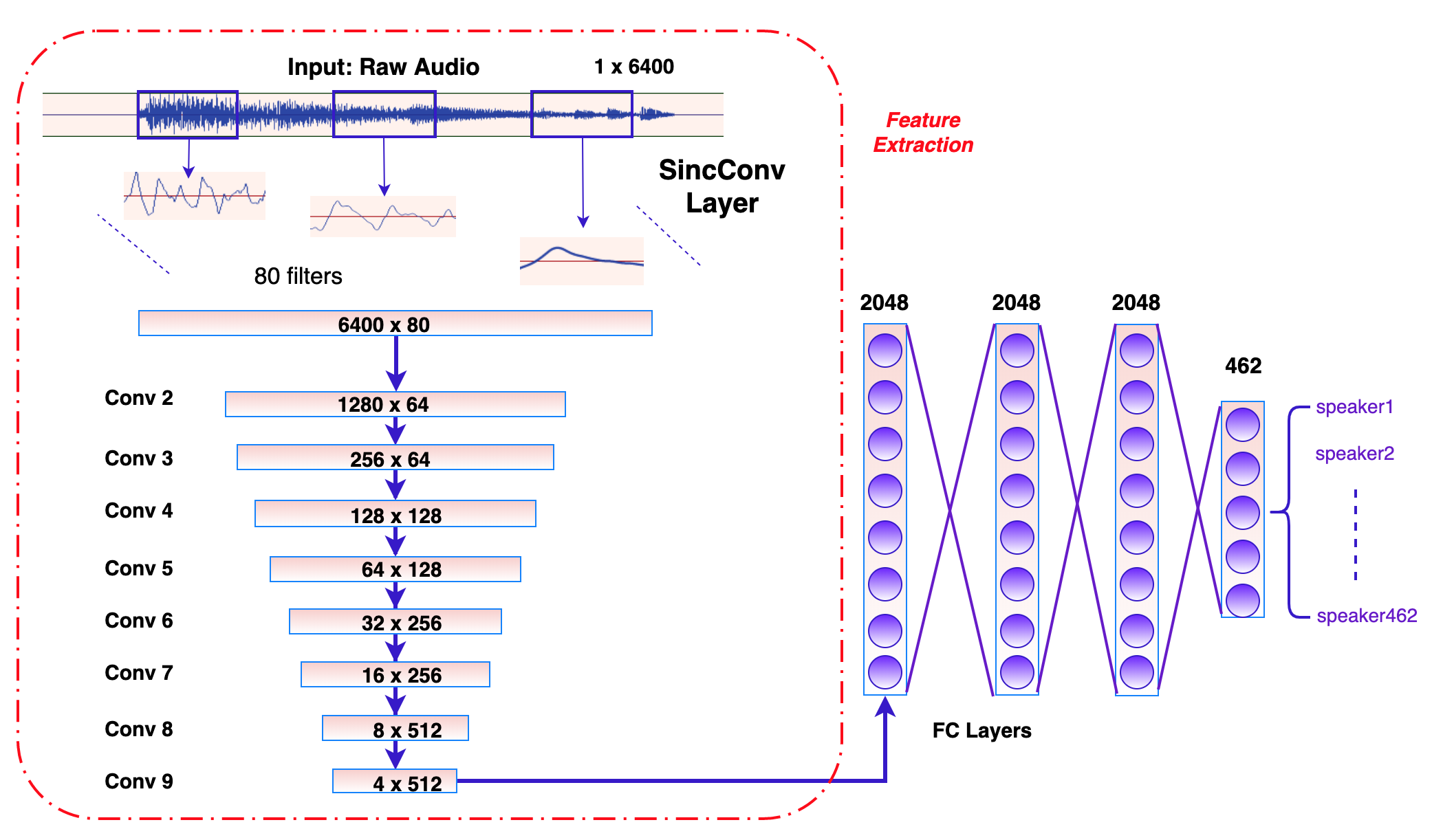


图 5.1 基于SCNN网络的说话人识别系统示意图(TIMIT数据集[42]下)

* 1. 实验设置及结果分析对比
     1. 实验设置

1. 数据集

本文提出的说话人识别系统采用的数据集是TIMIT数据集[42]。

TIMIT数据集是为声学语音研究以及自动语音识别系统的开发和评估提供的语音数据集。TIMIT包含由八种主要的美式英语方言组成的630个扬声器的宽带录音，每个方言都有10个丰富的语料句子。TIMIT语料库包括音素级别的语料标注以及每个话语的16-bit，16kHz语音波形文件。TIMIT数据集是麻省理工学院（MIT），SRI国际（SRI）和德州仪器公司（TI）共同完成的，被广泛地应用于说话人识别、语音识别等任务。

2. 系统结构及参数设置

实验在Pytorch框架下进行，与4.3节提出的系统参数相同，在第一层SincConv卷积层，采用的滤波器数目为80，卷积核大小为81个单位长度(约5ms)，卷积步长为1个单位长度。后端CNN网络采用卷积核大小为3的卷积层，每一层之间采用了批标准化(BN)，采用Leaky ReLU激活函数。

对于说话人识别任务，每个语音句子的波形被分成200ms（重叠10ms）的片段[29]，这些片段输入到该特征提取网络中，连接由2048个神经元组成的三个全连接层并用批量归一标准化（BN），最后通过Softmax获得帧级别分类结果。该系统使用RMSprop优化器进行训练，学习率为，，以及batch size设置为64。

* + 1. 实验结果分析及对比

表 5.1展示了说话人识别系统下SCNN网络性能的测试结果，统计其分类错误率。同时，测试了MFCC、FBANK等人工特征和标准CNN结构的错误率作为对比。实验结果表明，基于SCNN网络的说话人识别系统效果显著。

|  |  |
| --- | --- |
| 表 5.1 Classification Error Rate of TIMIT Dataset | |
| Model | CRE (%) |
| DNN-MFCC | 0.99 |
| CNN-FBANK | 0.86 |
| CNN-Raw | 1.65 |
| **SCNN (ours)** | **0.87** |

为了进一步体现该声学特征模型(SCNN)的特性，绘制了其在TIMIT数据集上的累积频域响应图，如图 5.2。可以观察到，频域范围集中在5000Hz以下，这与人发音在30Hz～3000Hz范围内的事实[40]相符。其次，该累积频域响应曲线存在一些细节信息，可以看到存在几处明显的峰谷[29]。第1个峰值对应于音高领域，男性的平均音高为133Hz，女性的平均音高为234Hz。第2个峰值(大约位于500 Hz)主要捕获第一共振峰，实际上，英语元音的平均值约是500 Hz。第3个峰值（范围从900到1400Hz）捕获一些重要的第二共振峰，例如元音a的第二共振峰，其平均值位于1100Hz。这些信息侧面反映了SCNN网络成功学习到了语音的特征。

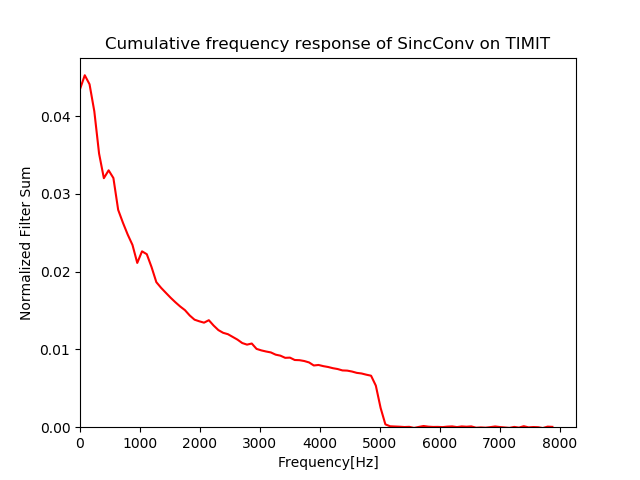


图 5.2 SincConv卷积层在TIMIT数据集下学习到的累积频域响应图。

* 1. 本章小结

本章将SCNN网络应用于说话人识别系统，并取得了较好的效果。主要介绍了说话人识别任务，基于SCNN模型的系统设计和相关实验情况。

1. 总结与展望
   1. 论文工作总结

随着人工智能的发展，和现实生活中对于人机交互的需求增加，机器听觉越来越被学术界和工业界所重视。绝大多数机器听觉系统都需要适当的特征提取技术，一个好的声学特征往往对于许多声学处理任务起着极其重要的作用。

本文首先介绍了课题的研究背景，针对传统的声学特征提取方法存在的问题和局限性进行讨论，如何获得一个相对普适的特征提取方法而不囿于所研究的问题成为本文研究的核心。

接着，论文介绍了一些主流的方法，分析比较后提出一种基于CNN的深度声学特征提取方法，即SCNN网络。第三章详细介绍了SCNN网络的设计思路，具体结构以及优势。

第四章和第五章介绍了基于此SCNN网络的多任务系统，包括的任务有声学事件检测，声学场景分类以及说话人识别。然后通过一系列实验，验证了SCNN网络对于各个任务处理的有效性。

本文提出一种可以在多任务下提取深度声学特征的网络模型，可以提取出声音信号的一种时频特征表示，具有一定的可解释性。同时，该网络模型已被实验证明适用于声学事件检测、声学场景分类及说话人识别任务。

在进行本文工作的同时，笔者收获颇多。从开始的知识学习阶段到实验测试阶段，再到论文撰写阶段。笔者学习了很多深度学习相关知识、机器听觉和信号处理相关知识以及Linux系统和Pytorch框架的使用等等。在这一过程中，通过阅读大量的文献，丰富了笔者对深度学习领域的认识，提高了笔者的科研素养和专业认同感。撰写论文的过程，增强了笔者的总结能力和语言表达能力，提高了笔者的综合素质。

* 1. 未来工作展望

关于深度声学特征提取，未来还需要开展多项工作。一些重要的问题尚未解决，经过笔者在本科生毕业设计阶段的研究和总结，未来的工作概括为以下几点：

1.尽管本文提出的深度声学特征模型于一些分类任务上效果尚可，但是声学处理任务种类繁多，使得其在更多的任务上提供帮助尚需实验验证；

2.本文提出的深度声学特征模型虽然在不同的任务上均优于常规方法，但是未通过实验测试其迁移能力，即在某一任务上训练的模型是否会对其他任务形成促进，这是一个值得研究的工作；

3.目前该深度声学特征模型的输入具有一定局限性，即其根据不同任务需要设计不同的输入长度，后续工作可能将尝试解决此问题；

4.传统的声学特征提取方法可以通过组合的方式互相促进，本文提出的方法与其他方法的结合是否有效，将是一个有意义的工作。

插图索引

[图 2.1声学特征提取标准流程图 4](#_Toc10026013)

[图 2.2在餐厅录制的声音信号的不同时域和时频域表示图：0.5秒时有人正在清理喉咙，2秒时有一些餐具噪音[13]。(a)时域波形。(b)线性频谱图。(c)Mel谱图。(d)Constant-Q谱图。 6](#_Toc10026014)

[图 2.3特征设计过程示意图 8](#_Toc10026015)

[图 2.4梅尔滤波器频带(a) MFCC分解(b) 10](#_Toc10026016)

[图 3.1基于分类任务下的SCNN网络结构图 17](#_Toc10026017)

[图 3.2 SCNN卷积层间结构图 18](#_Toc10026018)

[图 3.3 标准卷积层示意图 18](#_Toc10026019)

[图 3.4 数据集ESC-50下SCNN模型学习到所有80个滤波器的频域响应图，x轴为频率(Hz)，y轴为滤波器数目(每8个显示)，z轴为频率响应。 21](#_Toc10026020)

[图 3.5 数据集ESC-50上一组数据的时域波形(a)和对数梅尔频谱图(b) 21](#_Toc10026021)

[图 3.6 数据集ESC-50下SCNN模型学习到的3组滤波器的时域和频域响应图。a图为第10个滤波器的时域图，b图为第10个滤波器的频域图(可见频带范围为253.71Hz～283.63Hz)，c图为第50个滤波器的时域图，d图为第50个滤波器的频域图(可见频带范围为2609.52～2714.61Hz)，e图为第75个滤波器的时域图，f图为第75个滤波器的频域图(可见频带范围为6538.36Hz～6768.8Hz)。 22](#_Toc10026022)

[图 3.7 数据集ESC-50下CNN模型学习到的3组滤波器的时域和频域响应图。a图为第10个滤波器的时域图，b图为第10个滤波器的频域图，c图为第50个滤波器的时域图，d图为第50个滤波器的频域图，e图为第75个滤波器的时域图，f图为第75个滤波器的频域图。 23](#_Toc10026023)

[图 3.8 数据集ESC-50上一组数据(见图 3.5)分别通过训练好的SCNN网络模型的第1个、第41个和第71个滤波器后得到的时域波形(a,c,e)和对数梅尔频谱图(b,d,f) 24](#_Toc10026024)

[图 3.9 数据集ESC-50[35]上一组数据(见图 3.5)分别通过训练好的CNN网络模型的第1个、第41个和第71个滤波器后得到的时域波形(a,c,e)和对数梅尔频谱图(b,d,f) 25](#_Toc10026025)

[图 4.1 基于SCNN网络的声学事件检测系统示意图(ESC-50数据集[35]下) 30](#_Toc10026026)

[图 4.2 基于SCNN网络的声学场景分类系统示意图(TUT Urban Acoustic Scenes 2018，Development dataset数据集[36]下) 30](#_Toc10026027)

[图 4.3 训练阶段Loss情况 34](#_Toc10026028)

[图 4.4 训练阶段Accuracy情况 34](#_Toc10026029)

[图 4.5 ESC-50数据集上混淆矩阵 35](#_Toc10026030)

[图 4.6 DCASE Task1数据集上混淆矩阵 35](#_Toc10026031)

[图 4.7 SincConv卷积层在ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)下学习到的累积频域响应图。 36](#_Toc10026032)

[图 4.8 随着训练的进行，在ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)下学习到累积频域响应图变化过程。x轴为频率(Hz)，y轴为Epoch(每50个显示)，z轴为频率响应。 37](#_Toc10026033)

[图 4.9 ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)中随机两组数据的对数梅尔频谱图 37](#_Toc10026034)

[图 5.1 基于SCNN网络的说话人识别系统示意图(TIMIT数据集[42]下) 40](#_Toc10026035)

[图 5.2 SincConv卷积层在TIMIT数据集下学习到的累积频域响应图。 42](#_Toc10026036)

表格索引

[表 3.1 几种主流深度特征提取方案的对比 16](#_Toc10026037)

[表 4.1 Comparison of Accuracy of ESC-50 Dataset 33](#_Toc10026038)

[表 4.2 Comparison of Accuracy of DCASE2018 Task1 34](#_Toc10026039)

[表 5.1 Classification Error Rate of TIMIT Dataset 41](#_Toc10026040)

参考文献

1. 张雪英. 数字语音处理及MATLAB仿真[M]. 电子工业出版社, 2010.
2. 李宏松, 苏健民, 黄英来,等. 基于声音信号的特征提取方法的研究[J]. 信息技术, 2006, 30(1):91-94.
3. LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436.
4. Lyon R F. Machine hearing: An emerging field [exploratory dsp][J]. Ieee signal processing magazine, 2010, 27(5): 131-139.
5. Alías F, Socoró J, Sevillano X. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds[J]. Applied Sciences, 2016, 6(5): 143.
6. Russell S J, Norvig P. Artificial intelligence: a modern approach[M]. Malaysia; Pearson Education Limited,, 2016.
7. Bakhtin M, Ghāsemipour G. The problem of speech genres[J]. Literary Criticism, 2011, 4(15): 114-136.
8. Downie J S. Music information retrieval[J]. Annual review of information science and technology, 2003, 37(1): 295-340.
9. Zhuang X, Zhou X, Hasegawa-Johnson M A, et al. Real-world acoustic event detection[J]. Pattern Recognition Letters, 2010, 31(12): 1543-1551.
10. Barchiesi D, Giannoulis D, Stowell D, et al. Acoustic scene classification: Classifying environments from the sounds they produce[J]. IEEE Signal Processing Magazine, 2015, 32(3): 16-34.
11. Seide F, Li G, Chen X, et al. Feature engineering in context-dependent deep neural networks for conversational speech transcription[C]//2011 IEEE Workshop on Automatic Speech Recognition & Understanding. IEEE, 2011: 24-29.
12. Computational analysis of sound scenes and events[M]. Heidelberg: Springer, 2018.
13. Mesaros A, Heittola T, Virtanen T. TUT database for acoustic scene classification and sound event detection[C]//2016 24th European Signal Processing Conference (EUSIPCO). IEEE, 2016: 1128-1132.
14. Fletcher H. Auditory patterns[J]. Reviews of modern physics, 1940, 12(1): 47.
15. Glasberg B R, Moore B C J. Derivation of auditory filter shapes from notched-noise data[J]. Hearing research, 1990, 47(1-2): 103-138.
16. Patterson R D, Robinson K E N, Holdsworth J, et al. Complex sounds and auditory images[M]//Auditory physiology and perception. Pergamon, 1992: 429-446.
17. O'shaughnessy D. Speech communication: human and machine[M]. Universities press, 1987.
18. Brown J C. Calculation of a constant Q spectral transform[J]. The Journal of the Acoustical Society of America, 1991, 89(1): 425-434.
19. O'Shaughnessy D. Linear predictive coding[J]. IEEE potentials, 1988, 7(1): 29-32.
20. Davis S, 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): 357-366.
21. Phan H, Hertel L, Maass M, et al. Car-forest: Joint classification-regression decision forests for overlapping audio event detection[J]. arXiv preprint arXiv:1607.02306, 2016.
22. Rakotomamonjy A, Gasso G. Histogram of gradients of time–frequency representations for audio scene classification[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(1): 142-153.
23. Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
24. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
25. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
26. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.
27. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
28. Zhu B, Wang C, Liu F, et al. Learning environmental sounds with multi-scale convolutional neural network[C]//2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018: 1-8.
29. Ravanelli M, Bengio Y. Interpretable Convolutional Filters with SincNet[J]. arXiv preprint arXiv:1811.09725, 2018.
30. Li Y, Zhang X, Jin H, et al. Using multi-stream hierarchical deep neural network to extract deep audio feature for acoustic event detection[J]. Multimedia Tools and Applications, 2018, 77(1): 897-916.
31. Cummins N, Amiriparian S, Hagerer G, et al. An image-based deep spectrum feature representation for the recognition of emotional speech[C]//Proceedings of the 25th ACM international conference on Multimedia. ACM, 2017: 478-484.
32. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.
33. Rabiner L R, Schafer R W. Theory and applications of digital speech processing[M]. Upper Saddle River, NJ: Pearson, 2011.
34. Mitra S K, Kuo Y. Digital signal processing: a computer-based approach[M]. New York: McGraw-Hill, 2006.
35. Piczak K J. ESC: Dataset for environmental sound classification[C]//Proceedings of the 23rd ACM international conference on Multimedia. ACM, 2015: 1015-1018.
36. Mesaros A, Heittola T, Virtanen T. A multi-device dataset for urban acoustic scene classification[J]. arXiv preprint arXiv:1807.09840, 2018.
37. 韩纪庆. 声学事件检测技术的发展历程与研究进展[J]. 数据采集与处理, 2016, 31(2):231-241.
38. 姜洪臣, 梁伟, 张树武,等. 音频场景分类的音频特征提取和分析[C]// 第八届全国人机语音通讯学术会议论文集. 2005.
39. 王乃峰. 基于深层神经网络的音频特征提取及场景识别研究[D]. 哈尔滨工业大学, 2015.
40. 基于深度学习的说话人识别系统[D]. 西安邮电大学, 2018.
41. Hansen J H L, Hasan T. Speaker recognition by machines and humans: A tutorial review[J]. IEEE Signal processing magazine, 2015, 32(6): 74-99.
42. Zue V, Seneff S, Glass J. Speech database development at MIT: TIMIT and beyond[J]. Speech communication, 1990, 9(4): 351-356.
43. Palaz D, Collobert R. End-to-end acoustic modeling using convolutional neural networks for automatic speech recognition[R]. Idiap, 2016.
44. Muckenhirn H, Doss M M, Marcel S. On learning vocal tract system related speaker discriminative information from raw signal using CNNs[C]//Proc. Interspeech. 2018.
45. Palaz D, Collobert R. Analysis of cnn-based speech recognition system using raw speech as input[R]. Idiap, 2015.
46. Lee J, Park J, Kim K L, et al. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms[J]. arXiv preprint arXiv:1703.01789, 2017.
47. Meng A. Temporal feature integration for music organisation[D]. PhD thesis, Technical University of Denmark (DTU), 2006. URL http://www2. imm. dtu. dk/pubdb/p. php, 2006.

致谢

在本文工作的进展过程中，非常感谢我的指导老师高飞飞老师和北京大学邹月娴老师对我的帮助。每次我遇到问题的时候，老师总是不厌其烦地为我解答困惑。感谢ADSPLAB的师兄师姐的帮助和照顾，实验室给予我家一样的感觉。师兄师姐教会给我许多优秀的科研方法，分享给我许多有帮助的参考文献。在组会上，大家针对我的工作进行讨论，给我很重要的建议。

感谢我的父母对我的支持，在这段时间里十分关心我的情况，家是我的后盾。感谢在毕业论文工作中给予我帮助的所有亲朋好友，最后工作的顺利完成与这些感动密不可分。

声明

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

签 名：\_\_\_\_\_\_\_\_\_\_\_\_\_ 日 期：\_\_\_\_\_\_\_\_\_\_\_\_

附录A 外文资料的中文翻译

语音、音乐和环境声音的物理和感知特征提取技术综述

摘要：赋予机器与人类相似的感测能力是工程和计算机科学中的普遍追求。在追求使计算机感知周围环境的过程中，人们已经进行了巨大的努力来允许机器和计算机以类似人的方式获取，处理，分析和理解他们的环境。计算机感知人类声学环境的能力被称为机器听觉。为实现这一宏大的目标，音频信号的表示至关重要。在本文中，作者提供了最相关的音频特征提取技术的最新评论，该技术用于分析最常见的音频信号：语音，音乐和环境声音。除了重新审视经典的方法外，我们还涉及到新的分析领域以及受到新颖的生物知识启发。这些方法按照分类法进行描述，分类法根据其物理或感知基础对其进行组织，随后根据计算领域（时间，频率，小波，基于图像，倒谱或其他域）进行划分。这些方法的描述伴随着它们应用于机器听觉相关问题的最新例子。

关键词：音频特征提取; 机器听觉; 音频分析; 音乐; 语音; 环境声音

第1章 引言

人类感知的主要途径之一是听觉。因此，在寻求使计算机以类似人的方式感知其环境的过程中，从广义上感知声学环境是一项关键任务。然而，对于机器进行解码，空间中特定点的声学环境可能非常复杂，因为存在高度多样性的同时声源（来自天然或人为来源），或者由于许多其他原因，比如由于存在高背景噪声，或存在远距离声源等。

机器听觉是计算机像人类一样听到的能力，例如，通过区分语音与音乐和背景噪声，将两个前者拉出来进行特殊处理，根据它们的起源。此外，它还包括分析环境声音以识别声音事件的到达方向（例如，汽车路过）的能力，此外还检测在特定情况下哪些是通常的或不寻常的（例如，街道上的枪击），以及对诸如动作，事件，地点，乐器或扬声器等声学对象的识别。因此，理想的听力机器将面对各种各样的可听声音，并且应该能够成功地处理所有这些声音。

任何机器听觉系统都需要对输入的音频信号进行深入分析，旨在充分利用其特定的特性。该分析从提取音频信号的适当参数开始，该参数显示其最重要的特征，该过程通常被称为音频特征提取。

从逻辑上讲，从音频信号中提取正确的功能是保证机器听力应用成功的关键问题。实际上，提取的特征应提供参数化信号的紧凑但描述性的视觉，突出显示那些对完成手头任务最有用的信号特征，无论是检测，识别，分类，索引，检索还是识别。当然，取决于信号的性质（即，语音，音乐或环境声音）和目标应用，更有趣的是这些提取的特征从物理或感知的角度反映信号的特征。

本文介绍了应用于机器听觉的主要音频特征提取技术的最新情况。为了完整起见，作者在这项工作中加入了经典方法。此外，作者还介绍了音频特征提取技术的最新进展，以及它们在语音，音乐和环境声音分析中应用的新例子。值得注意的是，过去十年中引入的大多数最近开发的音频特征技术都需要定义超出经典领域的新分析方法（即时间，基于频率和倒频谱），例如在小波域，除了引入基于图像和多线性或非线性表示，以及受到生物知识启发的显著增加。

本文的结构如下。第2章描述了任何机器听觉系统的主要构成块，重点关注音频特征提取过程。此外，鉴于将信号的性质与提取的特征类型相关联的重要性，我们详细说明了机器听觉应用中涉及的三种最常见类型的信号的主要特征：语音，音乐和环境声音。第3章描述了遵循分类法来描述经典和最近定义的音频特征提取技术。第4章描述了基于音频信号物理特性的方法的基本原理和主要原理的描述，而第5章中解释了那些试图在某种程度上包括参数化过程中的感知的方法。 最后，第6章给出结论。

第2章 机器听觉

赋予机器能够感测其声学环境的问题通常通过面对特定的子问题来解决，例如特定类型的声音事件，场景或构图的检测，识别，分类，索引，检索或识别。其中，语音，音乐和环境声音构成了我们在机器听觉系统中的绝大多数声音类型。

在本章中，首先简要介绍机器听觉系统通用架构的主要目标和特征。然后，详细描述了那些系统处理的音频源的主要特征，即语音，音乐和环境声音。

2.1机器听觉系统的体系结构

无论解决的具体问题类型如何，底层系统的结构都可以通过图A-1中描述的通用和通用架构设计来描述。



图A-1 典型机器听觉处理流程图

在第一阶段，由麦克风捕获的连续音频流通过加窗被分段为较短的信号块。这是通过在理论上无限的输入信号样本流上滑动窗口函数来实现的，并通过将其转换为有限的样本块的连续序列而结束。由于窗口化，系统将能够在有限长度的样本块上操作。此外，取决于窗口函数的长度，可以假设典型的非静止音频信号在每个帧内是准静态的，从而便于后续的信号分析。

一旦输入的音频流被分段成有限长度的块，就从它们中的每一个中提取音频特征。 特征提取的目标是获得信号最显著的声学特征的紧凑表示，将N个样本长帧转换为K个标量系数（K << N），从而获得允许提高效率的数据压缩。为此，这些特征可以考虑在时域，频率等域中计算的信号内容的物理或感知影响。

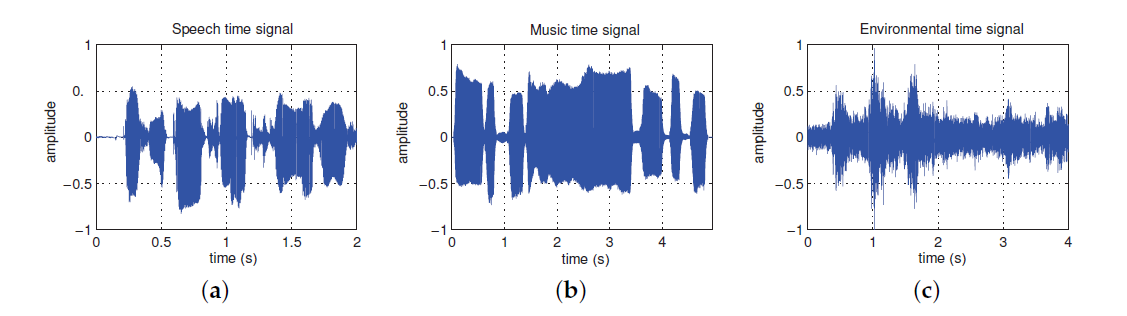
从这个意义上讲，对于执行某些类型的机器听觉任务，例如环境声音的识别或识别，已经发现建模音频信号的时间演变是至关重要的。为了保持该时间信息，可以将从若干后续信号帧中提取的特征合并为单个特征向量。应该注意的是，由于这个特征合并过程，特征向量获得了非常高的维度，这可能代表后续音频分析过程的障碍，具有所谓的维数爆炸问题。为了压缩特征向量，有时在特征提取技术之后是数据维数减少过程。

最后，必须对前一步骤中获得的特征向量进行音频分析任务。当然，音频分析是一种通用标签，试图包含解决手头特定机器听力应用所需的任何音频处理。当然，每个机器听力应用程序都要求根据应用程序特定的需求和要求设计音频分析块。

2.2 语音，音乐和环境声音的主要差异

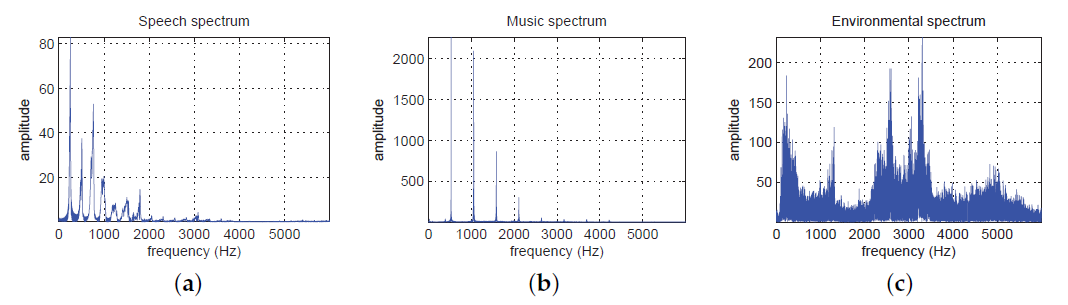
这里涉及要求机器听觉系统处理的音频输入，语音，音乐和环境声音呈现特定特征。可以在时域和频域以及信号的结构和语义中直接观察关键差异。然后，根据目标应用，可以根据物理或感知方法对这些差异进行参数化。

首先，音乐和语音信号呈现出一定的周期性，在时域中分析这些信号时可以观察到这种周期性，如图A-2所示。



图A-2 时域波形图：语音(a)，音乐(b)，环境声音(c)

其次，当在频域中分析时，通常可以确定环境声音的频谱（例如，经过的汽车的声音）的复杂性明显大于语音或音乐信号的复杂度，如图A-3所示。此外，可以观察到，语音和音乐信号通常在它们的光谱中呈现谐波结构，这是在环境声音中不常见的特性。



图A-3 标准化能量频谱图：语音(a)，音乐(b)，环境声音(c)

第三，语音和音乐声音都是由有限的声音单元字典组成：音素和音符。相反，环境声音的范围在理论上是无限的，因为环境中的任何发生的声音可以包括在该类别中（源自噪声，人造或自然声源）。

此外，这些类型的信号之间存在关键差异。在语音和音乐中，音素和音符被组合以便获得实际上传输特定语义或美学消息的有意义的序列。相反，环境声音上的序列不遵循任何规则或预定语法，尽管它们可以传达某种意义（例如，鸟鸣或蟋蟀声）。

第3章 音频特征分类与提取技术综述

存在无数种从机器听觉系统的音频输入中提取重要特征的方法。一方面，我们可以找到那些仅用于提取音频输入的物理特征的方法。这些提取技术在操作领域有所不同，从经典时间，频率或倒谱域到基于其他表示的特征的推导。具体地说，语音，音乐和环境声音通常具有丰富的时变特征，具有非常多样化的内容，其可以在该域中参数化。此外，这些音频信号的动态变化可以在变换域中呈现相关信息，例如，通过傅里叶变换，在倒谱或小波域中，或者从本征空间或甚至通过非线性表示，可以从中提取与例如频谱，谐度，线预测或相空间相关的特定特征。

在这项工作中，作者组织了对分层分类法的文献中最相关和最近的音频特征提取技术的回顾。作者通过区分基于物理的方法与具有感知基础的方法对技术进行分类，然后根据参数化的区域对其进行分类：时间，频率，小波，基于图像，倒谱或其他域。

第4章 物理音频特征提取技术

本章描述了文献中报告的主要物理音频特征提取技术，根据先前定义的分类法进行分类。

4.1时域物理特征

时域特征的最重要特征是它们不需要对原始音频信号应用任何类型的变换，并且它们的计算直接在信号本身的样本上执行。这种音频特征提取方法构成了最基本和最经典的方法之一。时域物理音频特征可以分为以下类别：基于过零的特征，基于幅度的特征，基于功率的特征和基于节奏的特征。

以下描述了属于这些类别的最常用的时域特征。

4.1.1基于过零率的物理特征

这种物理特征基于对分析的音频输入的单速率变化的分析，下面是几个机器听觉应用中使用的简单但有效的参数化。

过零率（ZCR）：定义为音频信号波形在一秒间隔内穿过零幅度电平的次数，它提供了信号主频分量的粗略估计。基于该标准的特征已经应用于语音/音乐辨别，音乐分类，在音乐和环境声音识别中演唱语音检测，在嘈杂条件下的语音活动检测或基于音频的监视系统。

线性预测过零率（LP-ZCR）：该特征被定义为原始音频的ZCR与从线性预测滤波器获得的预测误差的ZCR之间的比率。其用途旨在区分显示不同相关程度的信号（例如，在浊音和清音之间）。

4.1.2基于幅度的功能

基于幅度的特征是基于对信号的时间包络的非常简单的分析。

幅度描述符（AD）：它允许区分具有不同信号包络的声音，例如用于区分动物声音。它基于通过自适应阈值（电平交叉计算）基于其高和低幅度来收集信号段的能量，持续时间和持续时间的变化。

MPEG-7音频波形（AW）：此特征是从下采样波形包络计算的，它被定义为非重叠分析时间窗函数的最大值和最小值。AW已被用作环境声音识别的一个特征。

微光：它计算波形幅度的周期间变化。这一特征已被普遍应用于研究病理性声音。然而，它也被应用于区分声乐和非声乐区域与歌曲中的音频，表征咆哮和尖叫的歌唱风格，原型，分类和创建音乐声音或改进说话人识别和验证等。

4.1.3基于功率的功能

以下描述了基于信号功率的最相关和最经典的时间音频特征。

短时能量：使用基于帧的过程，短时能量（STE）可以定义为每个信号帧的平均能量。STE可用于检测从清音到语音的转换。此功能也被用于音乐起始检测，语音识别，环境声音识别等应用和基于音频的监视系统。

体积：根据刘等人的工作。体积定义为帧内波形幅度的均方根（RMS）。它已被用于语音分段应用。

MPEG-7时间质心：它代表包含信号最大平均能量的时间瞬间，它被计算为信号包络上的时间平均值（并以秒为单位测量）。时间质心已被用作环境声音识别领域的音频特征。

MPEG-7对数攻击时间：它表征特定声音的攻击（例如，对于音乐声音，乐器可以产生平滑或突然的过渡）并且它被计算为从声音信号开始到声音信号的经过时间的对数。除了应用于音乐起始检测，对数攻击时间（LAT）已被用于环境声音识别。

4.1.4基于节律的物理特征

节奏代表音乐和语音的相关方面，但它在环境和人类活动相关的声音（例如，火车的声音，手指敲击等）中也可能是重要的，因为它表征声音事件的结构组织（变化 能量，音高，音色等沿着时间轴。以下段落描述了文献中最相关和经典的基于节奏的特征。

脉冲度量：这是一种使用长时间带通自相关来确定声音在5秒窗口中的节奏程度的度量。其计算基于在六个频带中找到输出包络的峰值及其进一步比较，当所有子带呈现规则图案时给出高值。此功能已用于语音和音乐识别。

脉冲清晰度：它是一个高级音乐维度，表达了在给定音乐作品中的容易程度，或者在该片段中的特定时刻，听众可以感知潜在的节奏或音节脉动。该特征已被用于发现与定性测量的相关性。

波段周期性：这是衡量音频信号中节奏或重复结构强度的指标。频带周期性在频带内定义，并且它是沿子带自相关函数的最大峰值的所有信号帧的平均值获得的。

节拍频谱/频谱图：它是基于时间变化和滞后时间的二维参数化，从而提供反映速度的时间变化的可解释表示。节拍频谱在节奏周期显示与信号节奏特性相匹配的相关峰值。节拍频谱可用于区分具有不同节奏模式的音乐（或整个音乐信号内的部分之间）。

循环节拍频谱：或简称CBS，它表示音乐信号的速度，它将信号基本周期的倍数组合在一个速度等级中。 因此，CBS给出了歌曲基本节拍周期的更紧凑的表示。 该特征已用于音频检索领域。

节拍跟踪器：这一特征是基于信号子带分解和在每个子带中应用梳状滤波器分析的算法方法得出的。节拍跟踪器在很大程度上模仿人类在音乐中跟踪节奏节拍的能力，并且不仅可以获得节奏，还可以计算节拍定时位置。

节拍直方图：它提供更一般的节奏视角，并总结音乐信号中出现的节拍速度。在这种情况下，小波变换用于分解八度音阶中的信号，以便在每个子带中执行最突出的周期性的后续累积，以产生所谓的搏动直方图。此功能已用于音乐流派分类。

4.2频域物理特性

频域上的音频特征构成了文献中报道的最大音频特征集。它们通常从短时傅里叶变换（STFT）变换获得或从自回归分析得到。一般而言，物理频域特征描述了信号频率内容的物理特性。

4.2.1基于自回归的频率特征

基于自动回归的特征源自信号的线性预测分析，其通常捕获语音信号的典型频谱优势（例如，共振峰）。

下面描述了基于信号自回归的最常用的物理频率特征。

1.线性预测系数：或简称LPC，此特征表示全极点滤波器，捕获语音信号的频谱包络（SE）（出现在声道中的共振峰或频谱共振），并广泛用于语音编码和识别应用程序。 LPC也已应用于音频分割和通用音频检索。

2.线谱频率：文献中也称为线谱对（LSP），线谱频率（LSF）是用于量化和插值目的的LPC参数的鲁棒表示。它们可以被计算为构成LPC多项式表示的回文和反回忆多项式的根阶段，其分别代表声门关闭和打开时的声道。由于其内在的稳健性，它们被广泛应用于各种分类问题，如说话者分割，乐器识别和语音/音乐歧视。

4.2.2基于STFT的频率特性

这种音频特征通常源自从STFT计算获得的信号频谱图。虽然属于此类别的一些特征是根据频谱图包络的分析计算的（例如，子带能量比，频谱通量，频谱斜率，频谱峰值或MPEG-7频谱包络，归一化频谱包络和立体声平移频谱特征）其他是从STFT阶段获得的（如群延迟功能和/或修改的群延迟功能）。

以下列表总结了最广泛使用的基于STFT的功能。

1.子带能量比：它通常被定义为沿预定义的一组频率子带的归一化信号能量的度量。从广义上讲，它粗略地描述了光谱的信号能量分布。关于分析的子带的数量和特性（例如，Mel标度，ad-hoc子带等）存在不同的近似。它已被用于音频分割和音乐分析应用和环境声音识别。

2.光谱通量：或简称为SF，此特征定义为帧到帧频谱幅度差矢量的2范数（参见Scheirer和Slaney [61]），它描述了频率能量分布的突然变化。声音可用于检测音符开始，或者更一般地说，可用于检测频谱分布的显着变化。它测量功率谱变化的速度，并可用于确定音频信号的音色。此功能已被用于语音/音乐辨别，乐器分类，音乐类型分类和环境声音识别。

3.MPEG-7频谱包络和归一化频谱包络：音频频谱包络（ASE）是对数频率功率谱，可用于生成原始音频信号的缩小频谱图，它是通过将一系列频带内的原始功率谱的能量相加而得到的。每个分贝尺度的光谱矢量用RMS能量包络进行归一化，从而产生称为归一化音频包络（NASE）的ASE的归一化对数功率版本。ASE特征已被用于音频事件分类，音乐类型分类和环境声音识别。

4.2.3亮度相关的物理频率特性

亮度是与高频和低频信号能量平衡密切相关的属性（当声音具有比低频率更高的声音时，声音是明亮的）。

1.光谱质心：简称SC，此功能描述光谱能量的重心。它可以被定义为信号帧幅度谱的第一个时刻（平均值的频率位置）。SC揭示了信号的主要频率，已应用于音乐起始检测，音乐分类，环境声音识别以及最近的音乐情绪分类。

2.光谱中心：此功能定义为信号频谱的中频，其中较低和较高的能量均衡。因此，是一种接近光谱质心的度量。它已被证明对音乐信号中的自动节奏跟踪很有用。

4.2.4与音调相关的物理频率特征

基频被定义为谐波静止音频信号的最低频率，其又可以被认为是音调声音。在音乐中，音调是根据音乐标准组织音阶音符的系统。以下段落描述了最广泛使用的与音调相关的特征。

1.基频：它也表示为F0。MPEG-7标准将音频基频特征定义为局部归一化光谱时间自相关函数的第一个峰值。

2.音高直方图：音调直方图不是使用非常特定的局部描述符，如基频，而是更紧凑地描述信号的音高内容。音高直方图已经被用于音乐类型分类。

3.音高曲线：此功能是音高的更精确表示，因为它考虑了真实乐器中产生的音调失调效果以及打击乐声音的音高表示。

4.谐波：此功能可用于区分音调或谐波（例如，鸟类，长笛等）和类似噪音的声音（例如，狗皮，军鼓等）。大多数传统的谐波特征要么使用脉冲序列来搜索F0倍数的峰值集合，要么使用自相关启发函数来找到信号的自重复。

5.谐波噪声比：谐波噪声比（HNR）计算为谐波部分的能量与信号其余部分的能量之间的关系，单位为分贝（dB）。虽然HNR一般用于分析病态声音，但它也被应用于一些音乐相关的应用，如咆哮和尖叫歌唱风格的表征。

6.MPEG-7频谱音色描述符：MPEG-7标准定义了一些与声音谐波结构密切相关的特征，适用于区分音乐声音。

4.2.5与色度相关的物理频率特征

色度与音高的感知有关，在某种意义上它是音调高度的补充。在音乐环境中，分开一个或多个八度音阶的两个音符具有相同的色度（例如，C4和C7音符），并且对人类听觉感知产生类似的效果。以下段落描述了与色度相关的频率特征，它们基本上是从直接物理方法计算出来的：

1.色度图：也称为基于色度的特征，色度图是一种基于频谱的能量表示，它考虑了八度音程中的12个音高等级（对应于音乐理论中的音高等级），它可以是从对数STFT计算。然后，它构成了一种非常紧凑的表示，适用于感知方法之后的音乐和谐波信号表示。

2.色度能量分布归一化统计：或简称CENS，该特征被设想用于音乐相似性匹配，并且已经显示出对节奏和音色变化的稳健性。因此，它可以用于识别给定乐曲的不同解释之间的相似性。

4.3基于小波的物理特征

小波是用于将给定函数或连续时间信号分成不同比例分量的数学函数。 小波变换（WT）具有优于传统傅立叶变换的优点，用于表示具有不连续性和尖峰的函数，以及用于精确地解构和重构有限，非周期或非平稳信号。

4.4图像域物理特征

这种特征提取方法基于音频信号的基于二维图像的联合。通常，其中一个维度对应于信号的频率视觉，而另一个维度对应于时间视图。频谱图图像特征，简称SIF，是包含一组技术的特征，这些技术专注于将图像处理领域的技术应用于时频表示（使用傅立叶，倒谱或其他类型的频率映射技术）。

4.5 倒谱域物理特征

倒谱特征是光谱的紧凑表示，并提供基于对数幅度的平滑近似。它们主要用于说话人识别和语音识别，但它们也被用于音频检索。

1.复数倒谱：定义为信号傅立叶变换的对数（具有展开相位）的逆傅里叶变换，并已用于语音信号的音高确定，也用于识别乐器。

2.线性预测倒谱系数：或简称LPCC。该特征被定义为线性预测频谱复数包络的对数幅度的逆傅立叶变换，并提供更强大和紧凑的表示，特别适用于自动语音识别和说话人识别，但也适用于歌手识别，音乐分类和环境声音识别。

第5章 感知音频特征提取技术

感知音频特征的概念为了找到基于人类感知来描述一般音频属性的方式。本章描述了主要的感知音频特征提取技术。

5.1时域感知功能

在时域感知特征的背景下，我们可以发现过零特征，基于感知自相关的特征和节奏模式。

5.1.1基于过零率的感知特征

1.过零峰值幅度（ZCPA）：设计用于嘈杂环境中的自动语音识别（ASR），显示出线性预测系数更好的结果。该特征是根据先前在若干心理声学缩放的子带中分解的信号的时域过零点计算的。在所有子带信号上的反零交叉长度的直方图上获得特征的最终表示。

2.间距同步过零峰值幅度（PS-ZCPA）：其设计旨在提高嘈杂环境下ASR的稳健性。原始方法基于听觉神经系统，因为它使用梅尔频率间隔滤波器组作为前端级。PS-ZCPA仅考虑反向过零长度，其峰值具有高于阈值的高度，该阈值是作为信号基音周期内的最高峰值的一部分而获得的。

5.1.2基于感知自相关的特征

自相关是时域中信号的自相似性的度量，具有对音频特征提取的各种应用。

1.自相关函数特征：或简称ACF，应用于环境声音分析并且适用于语音表示。为了计算ACF，首先根据音频信号计算自相关函数，然后通过一组与声学现象相关的基于感知的参数（信号响度，感知音高，感知音高强度和信号周期性）对该函数进行参数化。 。

2.窄带自相关函数特征：也称为NB-ACF，是从应用梅尔滤波器组之后计算的音频信号的自相关函数获得的（其基于Mel标度，由听众判断的音高的感知标度彼此相等）。这些功能已被证明可为室内和室外环境声音分类提供良好的性能。5.1.3节奏模式

这个特征是建立在特定响度感上的声学与调制频率的二维表示，它是通过对关键频带随时间的傅里叶分析并结合受人类启发的加权阶段而获得的。该特征已被证明在音乐相似性检索中是有用的。

5.2频域感知功能

可在感知频域上定义基于频率的特征。这种类型的特征基于考虑到人类听觉感知而测量的一些信号特性。

5.2.1基于调制的感知频率特征

基于调制的感知频率特征表示音频信号中存在的低频（例如，大约20Hz）调制内容，其产生幅度和频率变化。在包含节拍和节奏的音频信号（例如，音乐中的节奏模式，来自工业机械的音频信号，语音信号等）中容易观察到这些变化。该调制信息可以反映声音频率内容随时间的结构演变，并且可以针对每个频带单独测量。

5.2.2亮度相关的感知频率特征

感知频率特征的第二个子类型是那些旨在描述声音亮度的子类型。在这个子类别中，我们可以找到锐度，一种高频信号强度的度量，与音频亮度密切相关，它已被用于音频相似性分析。

5.2.3与音调相关的感知频率特征

音调是与谐波信号的主频率的主观感知密切相关的声音特性，并且它允许将类似噪声的声音与类似正弦曲线的声音区分开，尤其是那些频率与谐波相关的正弦声音。心理声学音调功能包含基于听觉的模型。这是一种模拟人体音调感知的测量方法，通过结合带通滤波，强调最相关的音调感知频带，使用特定的滤波器组模型（Gammatone）模仿耳蜗的频率选择性，并使用内部毛细胞模型，允许计算连续射击概率的自相关函数。

5.2.4响度相关的感知频率特征

本节总结了与音频信号响度相关的那些特征，这一概念被定义为声音强度的主观印象。复杂声音（例如，语音，音乐，噪声）的响度测量过程被描述为几个子带中的每个子带的响度指数（使用等响度轮廓）的总和。首先计算输入帧的功率谱然后通过减去（以dB为单位）听觉绝对阈值的近似值然后通过一组gammatone进行滤波来计算响度。滤波器和频率求和以产生每个听觉滤波器中的功率，其对应于作为频率的函数的内部激励。然后，这些激励被压缩，缩放并在滤波器上求和，以得到响度估计。

5.3基于小波的感知特征

在以下段落中，我们描述了最常用的基于小波的感知频率特征，它代表了基于小波的物理特征的扩展。

1.核心功率流方向系数（KPFOC）：使用一组2D内核来估计使用Gammatone滤波器组计算的听觉频谱图中每个点的功率流方向，获得了ASR前端，与先前的方法相比，对噪声和房间混响具有更强的鲁棒性，特别适用于小型词汇任务。

2.Mel频率离散小波系数：或简称MFDWC，通过将离散WT应用于从输入信号获得的Mel标度对数滤波器组能量来解释耳朵的感知响应。MFDWC最初被定义为改善语音识别问题，随后被应用于其他机器听力评估应用，如说话者验证/识别和基于音频的监控系统。

3.Gammatone小波特征：是在小波域中制定的音频特征的子类型，其解释了感知建模是Gammatone小波特征（GTW）。这些特征是通过Gammatone函数替换典型的母函数来获得的，它们对听觉系统进行建模。

5.4图像域感知功能

在本节中，包含心理声学或感知听觉模型的某些方面，基于图像的音频功能就得到扩展。

1.频谱图像特征：频谱图像特征来自前端参数化，这些参数化利用了心理声学模型。使用GHT从基于ASR的高斯混合模型和隐马尔可夫模型构建码本，以便训练人工神经网络，该人工神经网络学习用于在帧级音素识别应用中优化分类准确度的判别加权，MFCC用作前端参数化。

2.听觉图像模型：或简称AIM，这种特征提取技术包括模拟听觉谱分析，神经编码和时间积分的功能和生理模块，包括产生稳定听觉图像的周期性敏感时间积分的新形式。编码过程基于三阶段系统。简而言之，光谱分析阶段将声波转换为模型的基底膜运动（BMM）表示。神经编码阶段使BMM在水平上稳定并且锐化诸如元音共振峰的特征，以产生由听觉神经中的声音产生的神经活动模式的模拟。时间整合阶段稳定NAP中的重复结构并产生我们感知的模拟，称为听觉图像。

3.时间色度图像：此特征是音频信号的二维表示，用于绘制音频信号随时间的色度分布。该特征采用称为色度的色度概念的修改定义，其被定义为相隔n个八度音阶的所有音高的集合。结合从时间色度图像中提取局部指纹的指纹识别算法，所提出的特征允许提高音频复制检测和歌曲识别的准确性。

5.5倒谱域感知特征

基于感知滤波器组的倒谱特征基于基于倒频谱的参数的计算，该方法基于以下方法：首先，获得幅度的傅立叶变换的对数，或者使用具有一些可能的感知标准的特定滤波器组分解；第二，执行先前结果的傅里叶变换（或余弦变换）。

这种类型的特征包括众所周知的梅尔频率倒谱系数及其变体，其通常基于在最后的基于傅立叶的阶段之前使用不同的频率标度。以下段落描述了该领域最相关的功能。

1.Mel频率倒谱系数：也称为MFCC，主要用于语音识别领域，但也用于音频内容分类领域，因为他们的计算是基于第一阶段的基于感知的频率量表（人类听觉模型，其中受到频率Mel量表的启发）。在获得基于帧的傅立叶变换之后，对Mel级滤波器组的输出进行对数，并且最后通过离散余弦变换（DCT）对它们进行去相关。仅使用第一DCT系数（通常从8到13）来收集表示信号频谱包络的低频分量（主要与音色有关）的信息。MFCC也被用于音乐分类，歌手识别，环境声音分类，基于音频的监视系统。此外，在语音识别和说话者验证的背景下引入了MFCC的一些特定扩展，目的是在存在噪声的情况下获得更稳健的频谱表示。

2.格林伍德函数倒谱系数：建立在格林伍德的开创性工作基础之上，据说许多哺乳动物具有对数耳蜗频率响应，引入了Greenwood函数倒谱系数（GFCC），从物种特定的听力图测量中提取等响度曲线，作为音频特征提取，用于分析来自这些物种发声的环境声音。后来，这一特征也应用于多声道语音识别。

3.噪声强大的音频功能：或简称NRAF，这些功能包含基于三阶段过程的特定人类听觉模型（耳蜗中的第一阶段过滤，电活动中的机械位移转换、毛细胞中的对数压缩阶段、和使用去相关的减少阶段，模仿耳蜗核中的横向抑制网络）。

4.Gammatone倒谱系数：也称为GTCC，是一种基于Gammatone函数的滤波器组，可以准确地预测人类掩蔽数据。Gammatone滤波器组（使用4阶线性Gammatone滤波器）进行频率分析和音频再合成的有效实现。Gammatone滤波器最初设计用于模拟人类听觉光谱响应，因为它们在脉冲响应，幅度响应和滤波器带宽方面具有良好的近似性。类似Gammatone的特征也被用于音频处理，语音识别应用，水声事件检测远程监控应用，用于道路噪声源分类和计算听觉场景分析。

第6章 结论

本文提供了与机器听觉相关的最新最相关的音频特征提取技术，这些技术是为语音，音乐和环境声音的分析而开发的。为了给音频分析应用从业者提供一个独立的参考，本综述涵盖了可追溯到20世纪70年代的最基本和经典的音频特征提取方法，以及基于最新的音频特征推导贡献和新的计算领域和生物启发的范例。

为此，本文重新审视了经典音频特征提取技术作为参考，并通过考虑该研究领域的最新进展来扩展这些方法。除了使用在时间，频率和倒谱域上计算的特征扩展该评论之外，本文还描述了在小波和图像域上计算的特征提取技术，这些特征提取技术是从多线性或非线性参数化获得的，以及从特定表示（例如机器追踪）得到的特征提取技术。此外，值得注意的是，本文还描述了大量新颖的生物启发方法（例如，包括诸如Mel和Gammatone滤波器组的听觉模型，或者来自自相关函数或听觉图像模型的计算）。所描述的音频特征提取技术根据它们是否具有物理或感知基础而被分类。

此外，对所有已介绍的特征提取技术背后的主要概念和原理的描述已经考虑了三种主要类型的音频输入的特定特性：语音，音乐和环境声音。此外，本文通过介绍一些经典和最新的例子来说明这些技术在几个特定的机器听觉相关问题中的应用，例如，用于语音：分段，识别，说话者验证及识别或语言识别；用于音乐：注释，推荐，类型分类，乐器识别，歌曲识别或情绪分类；用于环境声音：识别，分类，基于音频的监视或计算听觉场景分析等。

最后，作者要指出的是，这项工作不是作为与音频分析相关的所有现有音频特征提取技术的全面集合，而是为了整理这一领域的文献中的最新方法研究。此外，作者期望在机器听觉中提出创新方法的新作品需要开发新颖的音频特征提取技术，这将扩展这项工作。

原文索引：

1. Alías F, Socoró J, Sevillano X. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds[J]. Applied Sciences, 2016, 6(5): 143.

综合论文训练记录表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学生姓名** | 王赫麟 | **学号** | 2015011522 | **班级** | 自51班 |
| **论文题目** | 深度声学特征学习及其应用 | | | | |
| **主要内容以及进度安排** | 主要内容：  声学特征提取是几乎所有声学处理任务的前提，同时所提取的声学特征也往往是决定声学处理系统性能的重要影响部分。目前，随着深度学习的发展，特征学习已成为越来越受欢迎的特征提取方法。主流的研究仍聚焦于通过深度学习的方法完成特定任务的声学特征学习，本文提出一种基于CNN的深度声学特征模型，称为SCNN网络，用以提取原始波形的深度声学信息，同时可以在声学事件检测、声学场景分析以及说话人识别等多个任务上有较好的表现。  进度安排：  2018.11-2018.12 查阅文献，做好调研，确定研究方案；  2018.12-2019.02 获取数据库，学习实验框架及相关知识和方法；  2019.02-2019.03 获取深度声学特征的有效表示模型；  2019.03-2019.05 设计基于此深度声学特征模型的声学事件检测、声学场景分类及说话人识别系统，进行实验测试其性能；  2019.05-2019.06 总结工作，撰写毕业论文。  **指导教师签字：**  **考核组组长签字：**  **年 月 日** | | | | |
| **中期考核意见** | **考核组组长签字：**  **年 月 日** | | | | |
| **指导教师评语** | **指导教师签字：**  **年 月 日** | | | | |
| **评阅教师评语** | **评阅教师签字：**  **年 月 日** | | | | |
| **答辩小组评语** | **答辩小组组长签字：**  **年 月 日** | | | | |

**总成绩：**

**教学负责人签字：**

**年 月 日**