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4月11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VGG）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。

本次毕业论文训练采用一定规模的音频数据库，进行深度声学特征模型的学习，并测试了基于此模型的声学事件检测任务和说话人识别任务。

**关键词：**深度学习； 深度声学特征； 声学事件检测； 说话人识别

ABSRTACT

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 no similar deep acoustic feature model for audio processing tasks.

This graduation thesis training uses a certain scale of audio database to learn the deep acoustic feature model, and tests the audio event detection task and speaker recognition task based on this model.

**Keywords:** deep learning; deep acoustic feature; audio event detection; speaker recognition

目录

[第1章 引言 1](#_Toc7871556)

[1.1研究背景 1](#_Toc7871557)

[1.1.1机器听觉 1](#_Toc7871558)

[1.1.2声学事件和声学场景分析 1](#_Toc7871559)

[1.1.3说话人识别 2](#_Toc7871560)

[1.1.4声学特征提取 3](#_Toc7871561)

[1.1.5深度神经网络 3](#_Toc7871562)

[1.2研究现状 4](#_Toc7871563)

[1.2.1人工声学特征提取方法概述 4](#_Toc7871564)

[1.2.1.1频谱图 4](#_Toc7871565)

[1.2.1.2 Log-Mel能量 4](#_Toc7871566)

[1.2.1.3 MFCC 4](#_Toc7871567)

[1.2.2 基于神经网络的深度声学特征提取概述 4](#_Toc7871568)

[1.2.2.1基于CNN的深度声学特征提取 4](#_Toc7871569)

[1.2.2.2基于DNN的深度声学特征提取 5](#_Toc7871570)

[1.2.2.3 基于RNN的深度声学特征提取 5](#_Toc7871571)

[1.3 毕业论文整体执行完成情况概述 5](#_Toc7871572)

[1.3.1 知识学习阶段 5](#_Toc7871573)

[1.3.2 方案设计阶段 5](#_Toc7871574)

[1.3.3 实验阶段 5](#_Toc7871575)

[1.3.4 系统调试、优化及性能分析阶段 5](#_Toc7871576)

[1.3.5 毕业论文工作总结阶段 5](#_Toc7871577)

[1.4 毕业论文组织结构安排 5](#_Toc7871578)

[第2章 研究方案 5](#_Toc7871579)

[2.1设计思路 5](#_Toc7871580)

[2.2 SincConv卷积层 7](#_Toc7871581)

[2.3后端网络设计 11](#_Toc7871582)

[2.4 相关方法概述 11](#_Toc7871583)

[2.4.1 激活函数 11](#_Toc7871584)

[2.4.2 对于梯度问题的处理 11](#_Toc7871585)

[2.4.3 损失函数 11](#_Toc7871586)

[2.4.4 优化器算法 11](#_Toc7871587)

[2.5 本章小结 12](#_Toc7871588)

[第3章 实验及相关配置 12](#_Toc7871589)

[3.1 实验数据集 12](#_Toc7871590)

[3.1.1 ESC-50数据集 12](#_Toc7871591)

[3.1.2 TUT Urban Acoustic Scenes 2018 dataset 12](#_Toc7871592)

[3.1.3 Urbansound8K 数据集 13](#_Toc7871593)

[3.1.4 TIMIT 数据集 13](#_Toc7871594)

[3.2 实验运行环境 14](#_Toc7871595)

[3.3 实验具体方法及参数设置 14](#_Toc7871596)

[3.4 本章总结 16](#_Toc7871597)

[第4章 实验结果分析 16](#_Toc7871598)

[4.1 实验过程分析 16](#_Toc7871599)

[4.2 模型性能测试 16](#_Toc7871600)

[4.2.1 模型性能评估指标 16](#_Toc7871601)

[4.2.2 模型性能 16](#_Toc7871602)

[4.3 模型分析 17](#_Toc7871603)

[4.3.1 频域响应图 17](#_Toc7871604)

[4.3.2 基于声学事件检测及声学场景分析任务的模型分析 17](#_Toc7871605)

[4.3.3 基于说话人识别任务的模型分析 18](#_Toc7871606)

[4.4 本章总结 18](#_Toc7871607)

[第5章 结束语 18](#_Toc7871608)

[5.1 本文主要内容梳理 18](#_Toc7871609)

[5.2 待解决问题及后续工作 18](#_Toc7871610)

[5.3 知识技能学习情况 18](#_Toc7871611)

[5.4 职业素养和工程理论的学习和培养 18](#_Toc7871612)

[5.5 毕业论文训练收获及体会 18](#_Toc7871613)

[参考文献 19](#_Toc7871614)

[致谢 19](#_Toc7871615)

第一章 绪论

1.1课题研究背景及意义

自从计算机被研发出来，实现人工智能(Artificial Intelligence, AI)一直是人类梦寐以求的事情。从声学的角度讲，人工智能追求的目标是让机器像人一样会听，会说并且会思考。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，语音和音乐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或声学场景来检测和识别。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音频特征提取。

传统研究表明，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（LPC）、LP倒谱系数（LPCC）、线谱对系数（LSP）、短时频谱、Mel频率倒谱系数（MFCC）等。还有结合时间和频率的特征，如频谱图。一般来说，这些特征都只包含了声学信号的部分信息。通常，为了充分表征声学信号，人们会尝试综合各种特征，并取得了一定的效果。然而，由于具体音频处理任务的限制和数学模型描述的局限性，人们往往未能充分利用所有的音频信息，而是根据具体问题进行特征的变换、取舍与组合等。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，就成了一个有意义而有挑战性的问题。

1.2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。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，Hinton等学者提出了深度置信网络(Deep Belief Networks, DBN)标志着神经网络研究的兴起。DBN是一种无监督的概率生成模型，通过逐层训练受限玻尔兹曼机(Restricted Boltzmann Machine, RBM)堆积得到。在训练好的DBN上添加一个和目标相关的输出层，例如分类任务的softmax输出层、回归任务的线性回归层，则可以构成一个深度神经网络(Deep Neural Networks, DNN)。基于DBN的方法使得DNN的训练优化变得容易，深层结构的模型优势开始展现。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。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。

1.3 研究内容和论文组织结构安排

声学特征提取是几乎所有音频处理任务的前提，同时所提取的声学特征也往往是决定音频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。

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，并将其与标准CNN网络结构进行对比分析。

第六章总结与展望：对全文进行总结，包括方法的选取以及实验结果分析，对于本文工作的创新点和不足之处进行总结，最后对未来的发展和方向进行展望。

第二章 声学特征研究概述

2.1 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。标准步骤如下图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)。

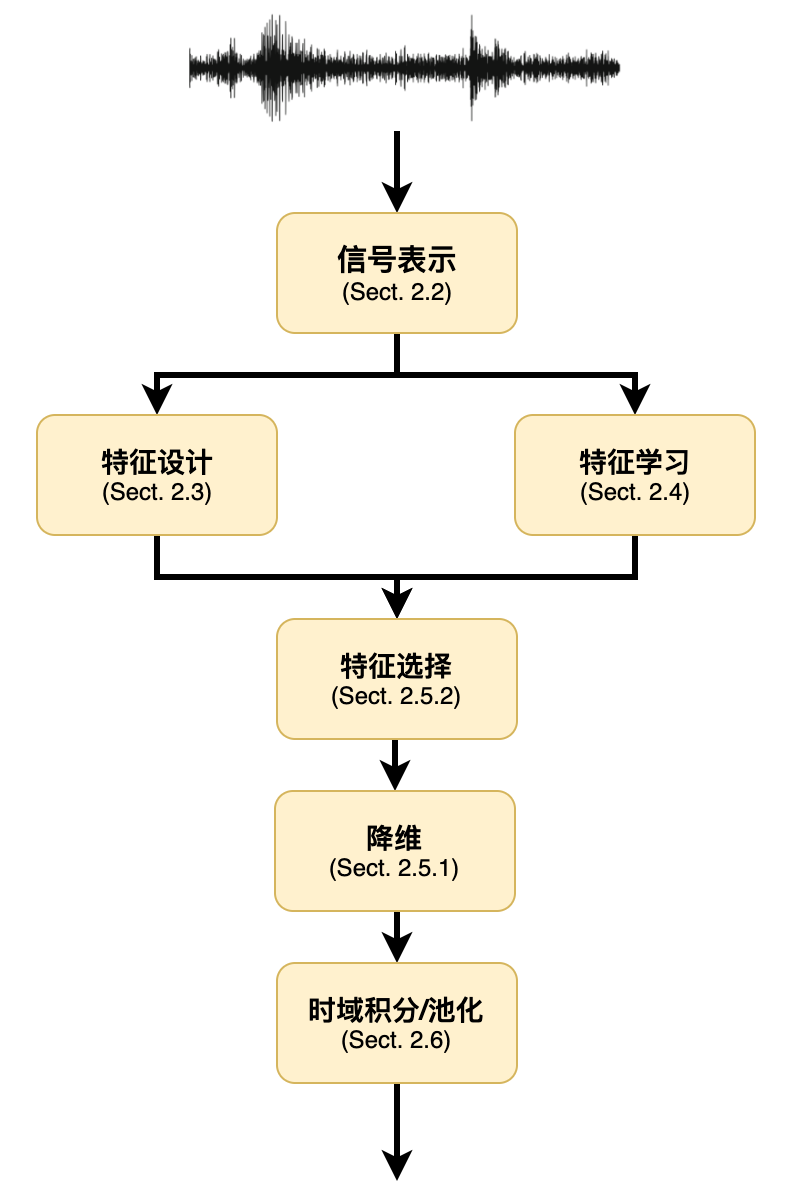


图1

2.2 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。

2.2.1 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：

* **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。
* **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。
* **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。）
* **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。

2.2.2 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：

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：

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的第个分量计算如下：

其中，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

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2b)。

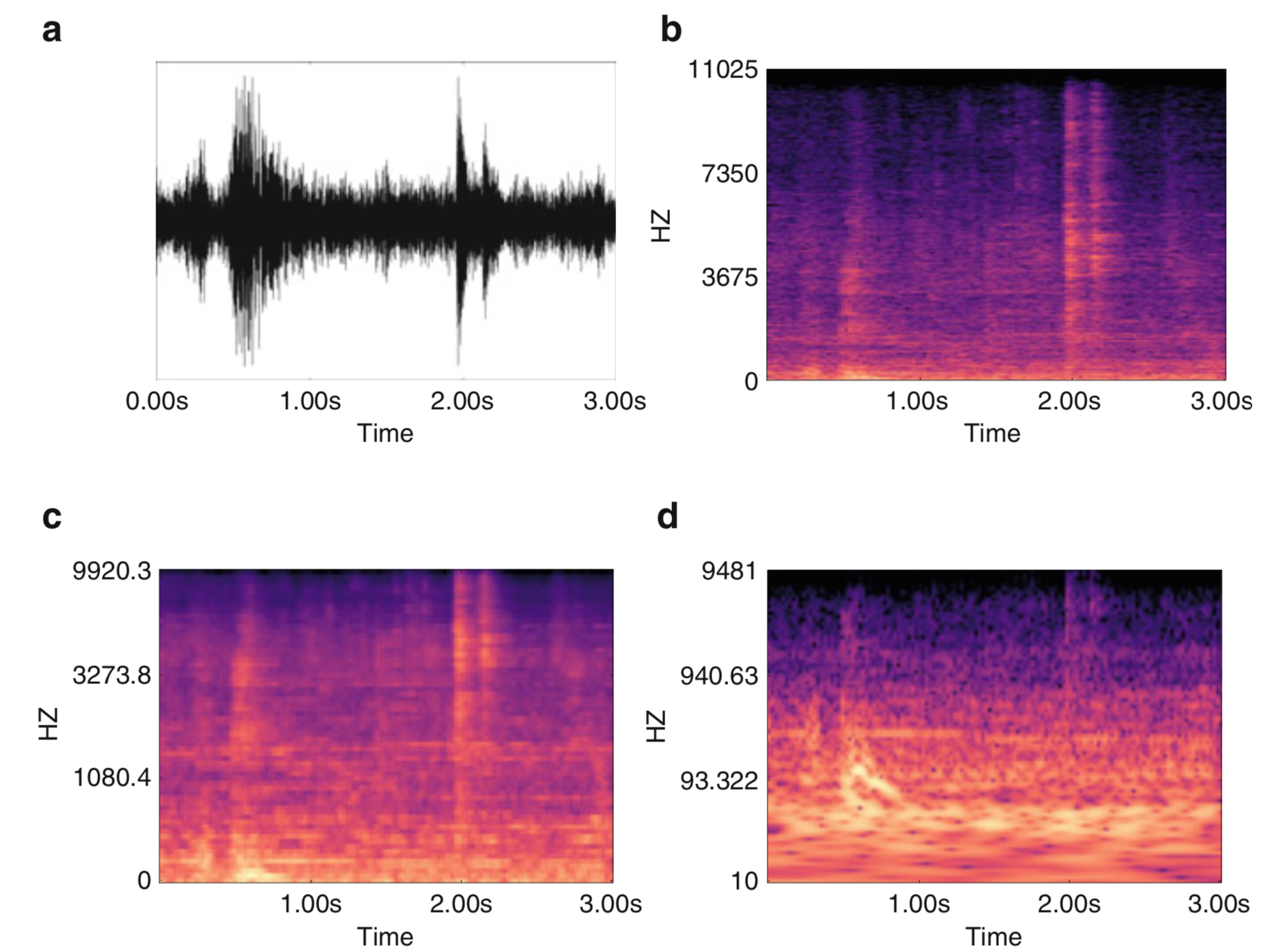


图2在餐厅录制的声音信号的不同时域和时频域表示：0.5秒时有人正在清理喉咙，2秒时有一些餐具噪音。(a)时域波形。(b)线性频谱图。(c)Mel谱图。(d)Constant-Q谱图

2.2.3 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。

* **关键频带**是由Fletcher提出的。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：

其中的单位是Hz。Bark标度是另一个依赖于关键频带概念的尺度，但它来源于不同的实验。

* **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：

其中，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，已经证明4阶的gammatone滤波器可以很好地逼近听觉滤波器。

* **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，存在几种解析表达式，由Fant给出了一种Mel标度和Hertz标度之间常用的关系:

• **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。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下式给出：

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。

2.3 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3)

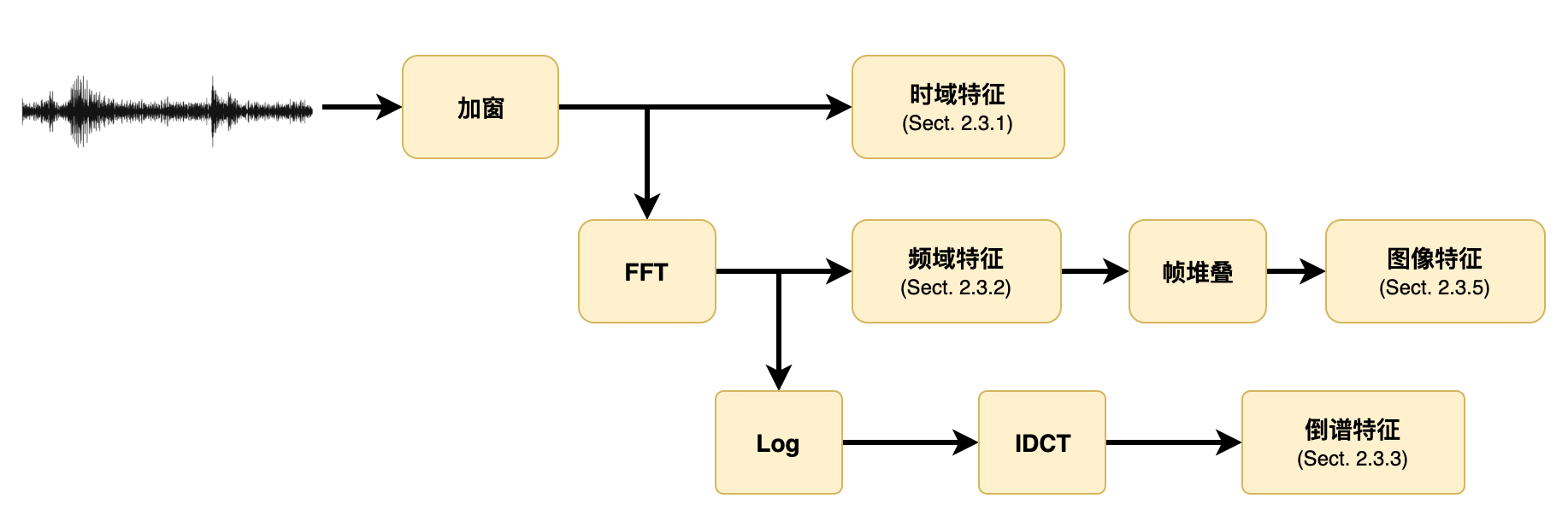


图3 特征设计过程

2.3.1 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。

* **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：
* **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，它由下式给出：

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。

* **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：

2.3.2 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。

* **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。
* **频谱包络**在概念上类似于时域包络，只不过是在频域中。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。
* **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。
* **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。
* **频谱斜率**测量频谱随频率降低的平均速率。
* **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。
* **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。
* **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。

2.3.3 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）。

* **梅尔频率倒谱系数(MFCC)**是最常用的倒谱系数。它们是在梅尔频带中对数能量的逆离散余弦变换获得的：

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。

* 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。

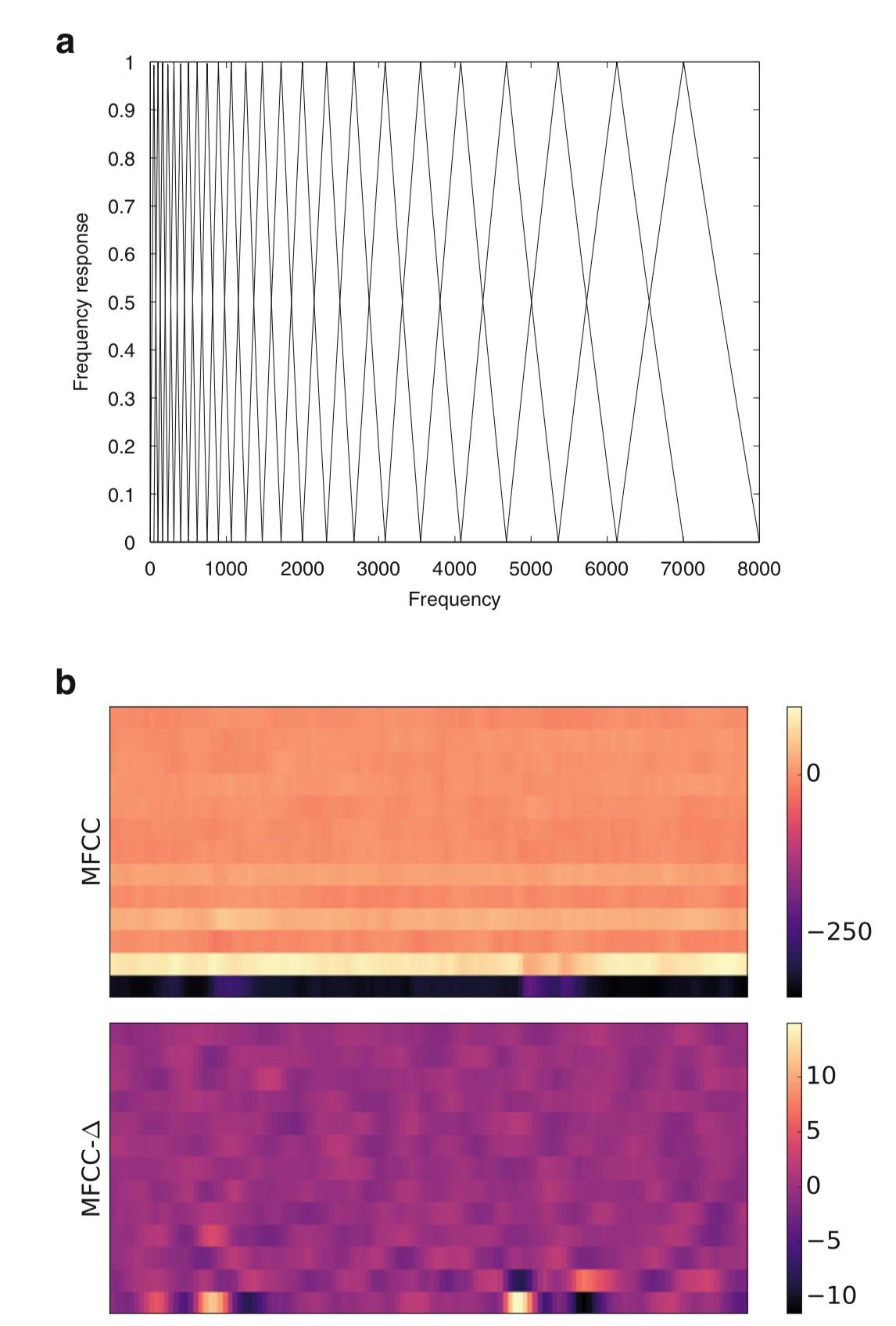


图4 梅尔滤波器频带(a)MFCC分解(b)

2.3.4 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：

* **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：

其中，是ERB频段的数目，每一个频段上的响度可以由下式近似得到：

其中，是第个频带的信号的能量。

* **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。
* **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：

2.3.5 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。

* **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。通常通过计算包含频谱图像中每个像素的梯度的梯度图像来提取。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。
* **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。
* **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。

2.4 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。

2.4.1 基于CNN的深度声学特征提取

自从2010年ImageNet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、VGGNet、Google Inception Net和ResNet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。由于卷积和池化计算的特殊性质，使得图像中的平移部分对于最后的特征向量没有影响。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，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；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。

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。

2.4.2 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。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。

2.5 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。

2.5.1 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。

2.5.2 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：

* 通过重复使用接近最优的搜索策略，可以避免暴力搜索。
* 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。

2.6 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。

2.6.1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。

2.6.2 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。

2.7 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。

第三章 基于CNN的深度声学特征模型设计

3.1 模型设计思路

3.1.1 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1通过对比几种主流的方法，确定本文的研究方案。

|  |  |  |  |
| --- | --- | --- | --- |
|  | **方案一** | **方案二** | **方案三** |
| **名称** | 根据语谱图提取深度特征 | 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。 |

表1 几种主流方案的对比

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。

3.1.2 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，第一层采用SincConv卷积层取代标准卷积层结构，下图5是基于分类任务下的网络结构图。卷积层之间采用Batch Normalization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，见图6。

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。

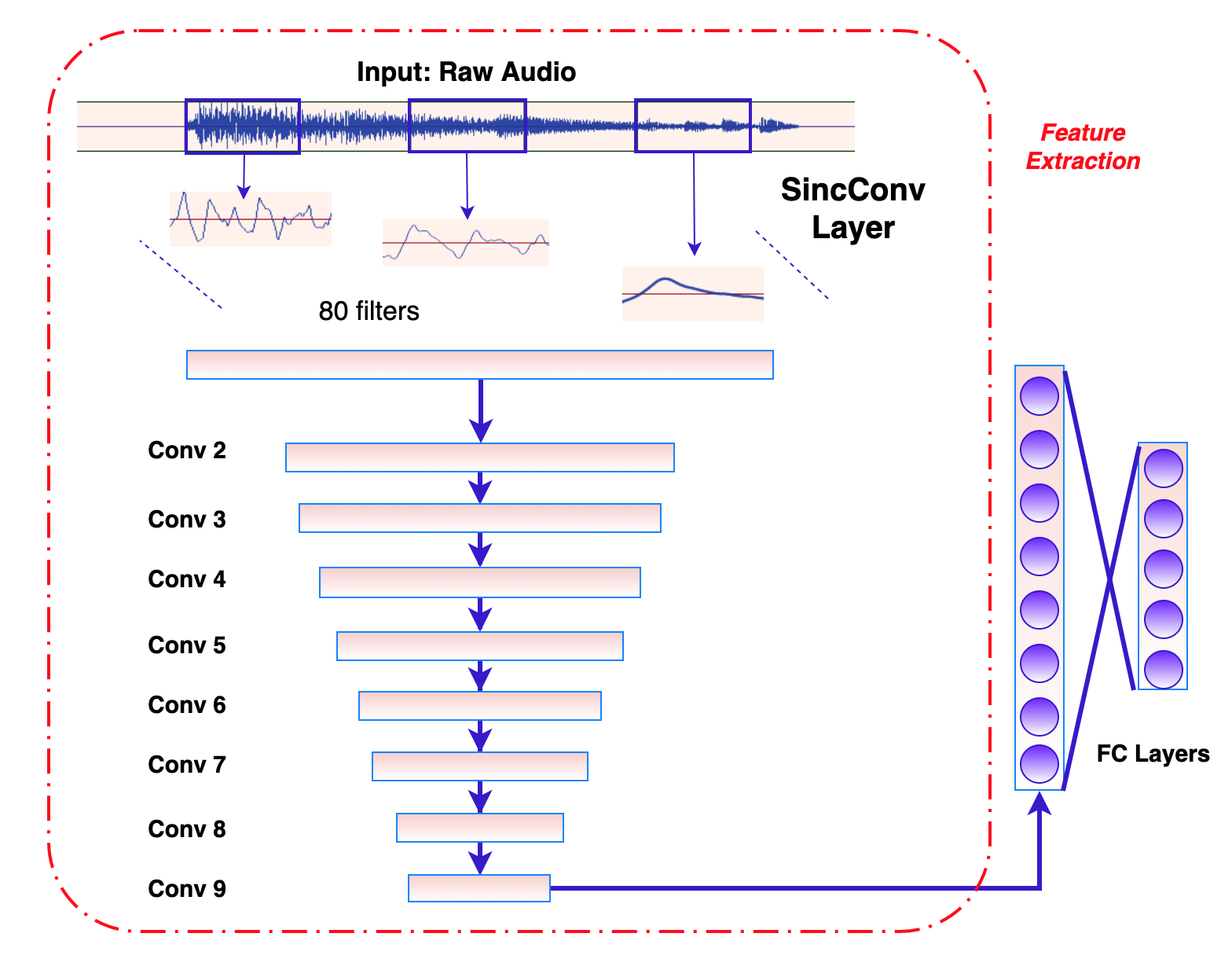


图5 基于分类任务下的SCNN网络结构图

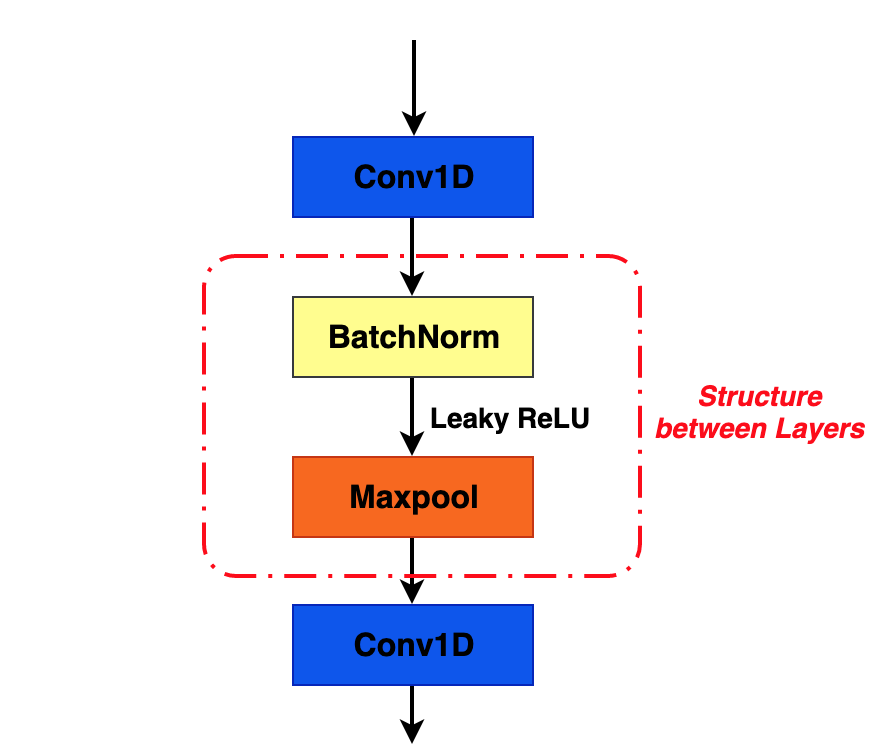


图6 SCNN卷积层间结构

3.2 SincConv 卷积层概述



图7 标准卷积层示意图

图7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)进行时序卷积，可以用下面公式表示卷积计算过程：

这里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：

这里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：

这里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），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，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：

SincConv卷积层中使用的是Hamming窗，定义如下：

Hamming窗特别适合实现高频选择性。但是，实验结果显示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程序与其它层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8为几个学习到的滤波器的时域和频域响应图示例。图10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，将以一层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9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模型的对比可以看出，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。

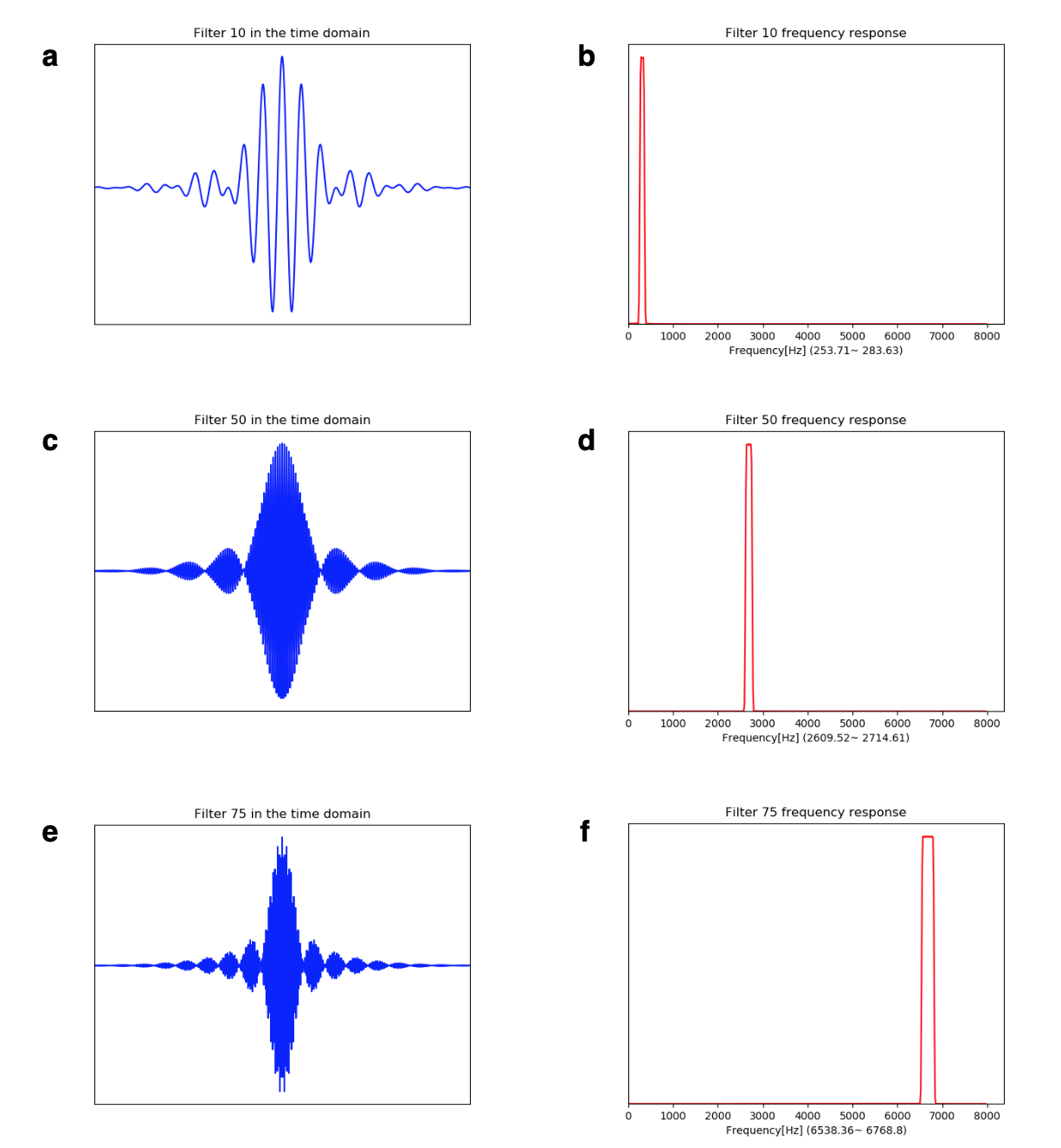


图8 数据集ESC50下SCNN模型学习到的3组滤波器的时域和频域响应图。a图为第10个滤波器的时域图，b图为第10个滤波器的频域图(可见频带范围为253.71Hz～283.63Hz)，c图为第50个滤波器的时域图，d图为第50个滤波器的频域图(可见频带范围为2609.52～2714.61Hz)，e图为第75个滤波器的时域图，f图为第75个滤波器的频域图(可见频带范围为6538.36Hz～6768.8Hz)。

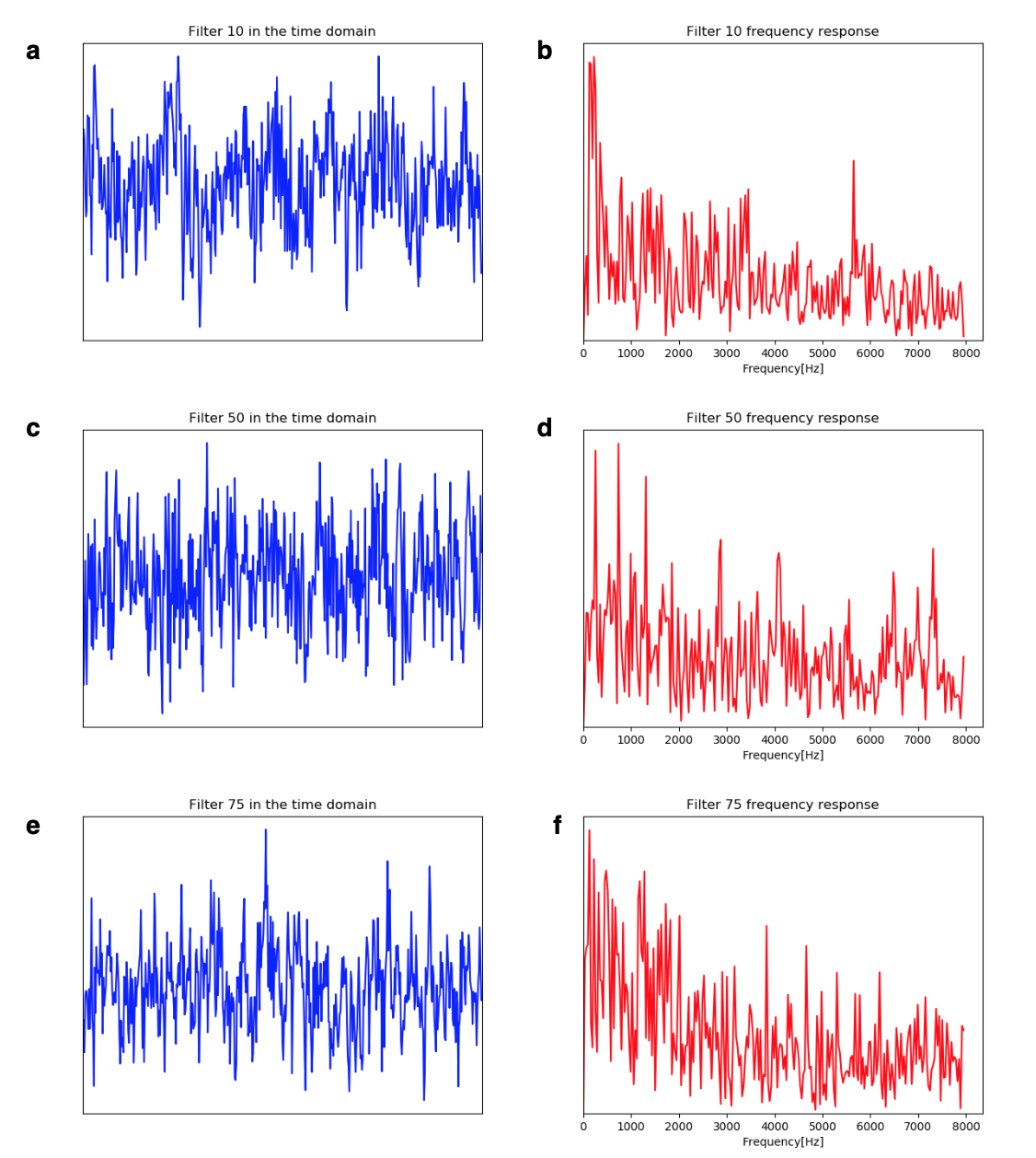


图9 数据集ESC50下CNN模型学习到的3组滤波器的时域和频域响应图。a图为第10个滤波器的时域图，b图为第10个滤波器的频域图，c图为第50个滤波器的时域图，d图为第50个滤波器的频域图，e图为第75个滤波器的时域图，f图为第75个滤波器的频域图。

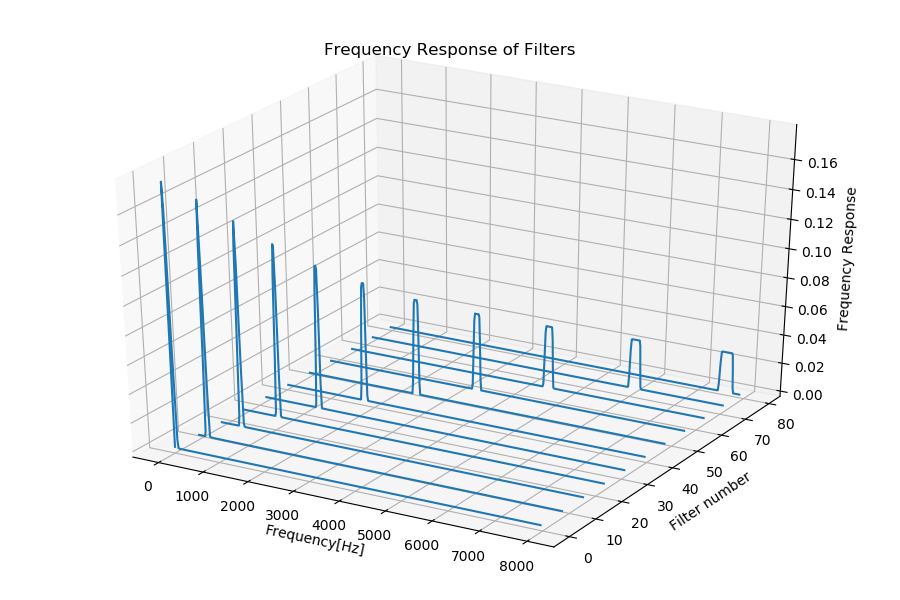


图10 数据集ESC50下SCNN模型学习到所有80个滤波器的频域响应图，x轴为频率(Hz)，y轴为滤波器数目(每8个显示)，z轴为频率响应。

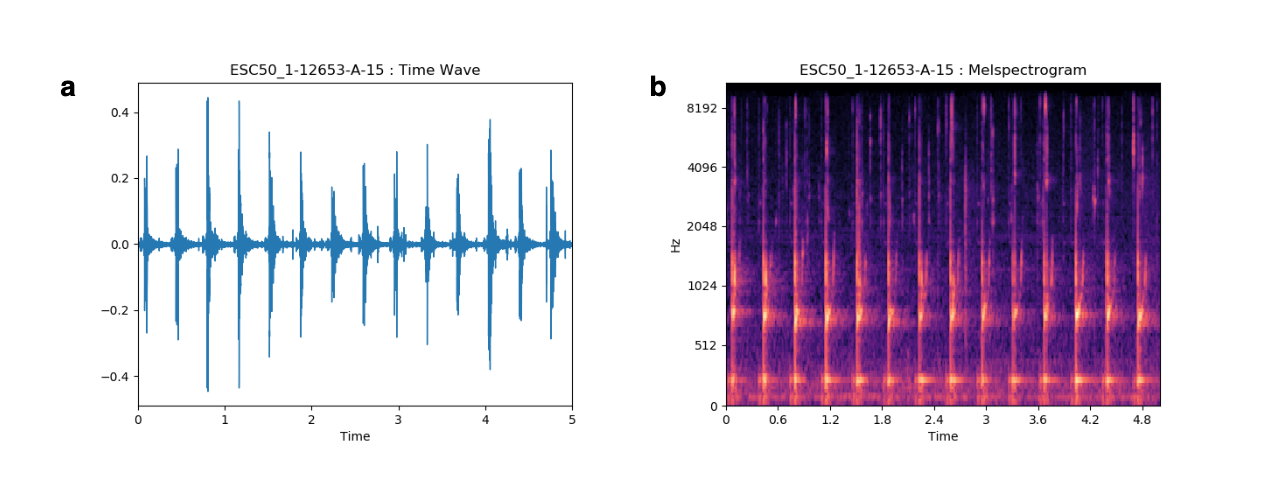


图11 数据集ESC50上一组数据的时域波形(a)和对数梅尔频谱图(b)

为了更明显地反映SincConv层的滤波效果，在一段音频数据上(图11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12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。

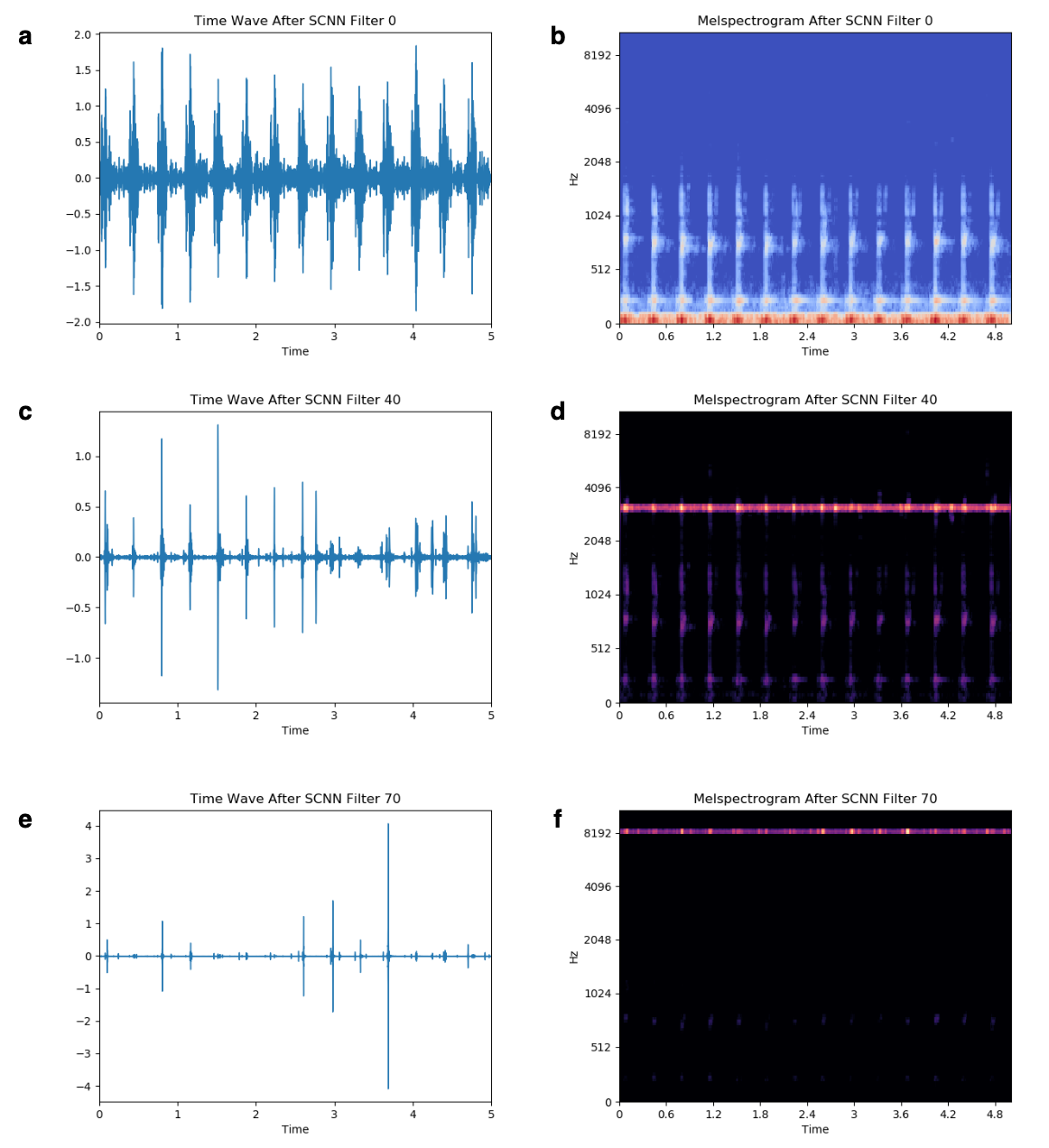


图12 数据集ESC50上一组数据(见图11)分别通过训练好的SCNN网络模型的第1个、第41个和第71个滤波器后得到的时域波形(a,c,e)和对数梅尔频谱图(b,d,f)

3.3 后端CNN网络设计概述

网络结构如图5，第2层至第9层CNN网络借鉴VGGNet的设计，采用卷积核大小为3的卷积层，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。

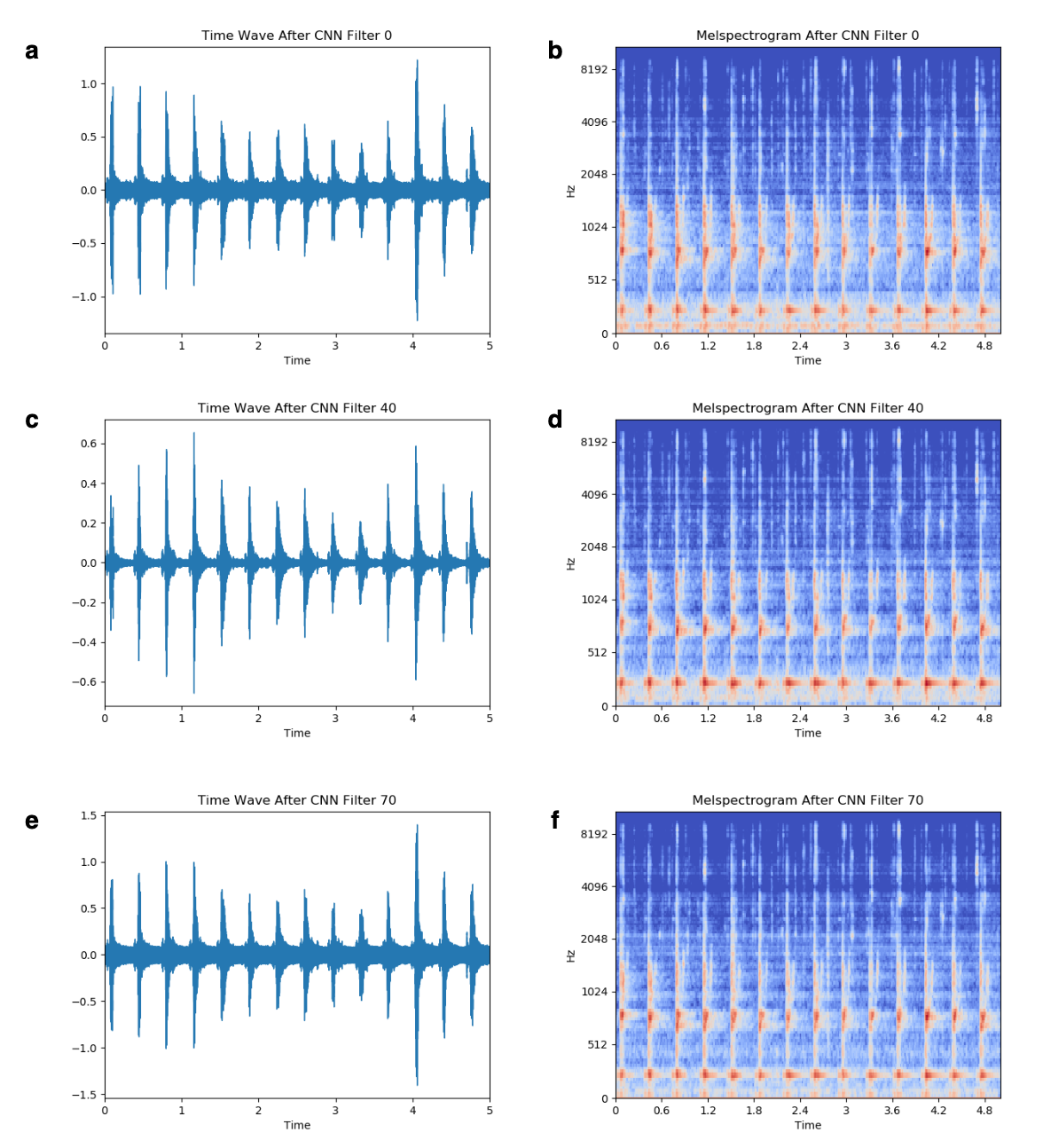


图13 数据集ESC50上一组数据(见图11)分别通过训练好的CNN网络模型的第1个、第41个和第71个滤波器后得到的时域波形(a,c,e)和对数梅尔频谱图(b,d,f)

3.4 相关方法介绍

3.4.1 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，相比于sigmoid函数和tanh函数，ReLU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:

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。

3.4.2 损失函数

SCNN网络采用交叉熵损失函数，其计算公式为：

其中，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，其计算方式如下：

3.4.3 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。

3.4.4 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5)使用Dropout方法防止过拟合。

3.5 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。

第四章 基于SCNN网络的声学事件检测和声学场景分类系统

4.1 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节介绍相关实验及结果分析。

4.2 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。

4.3 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14。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。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，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。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。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。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。

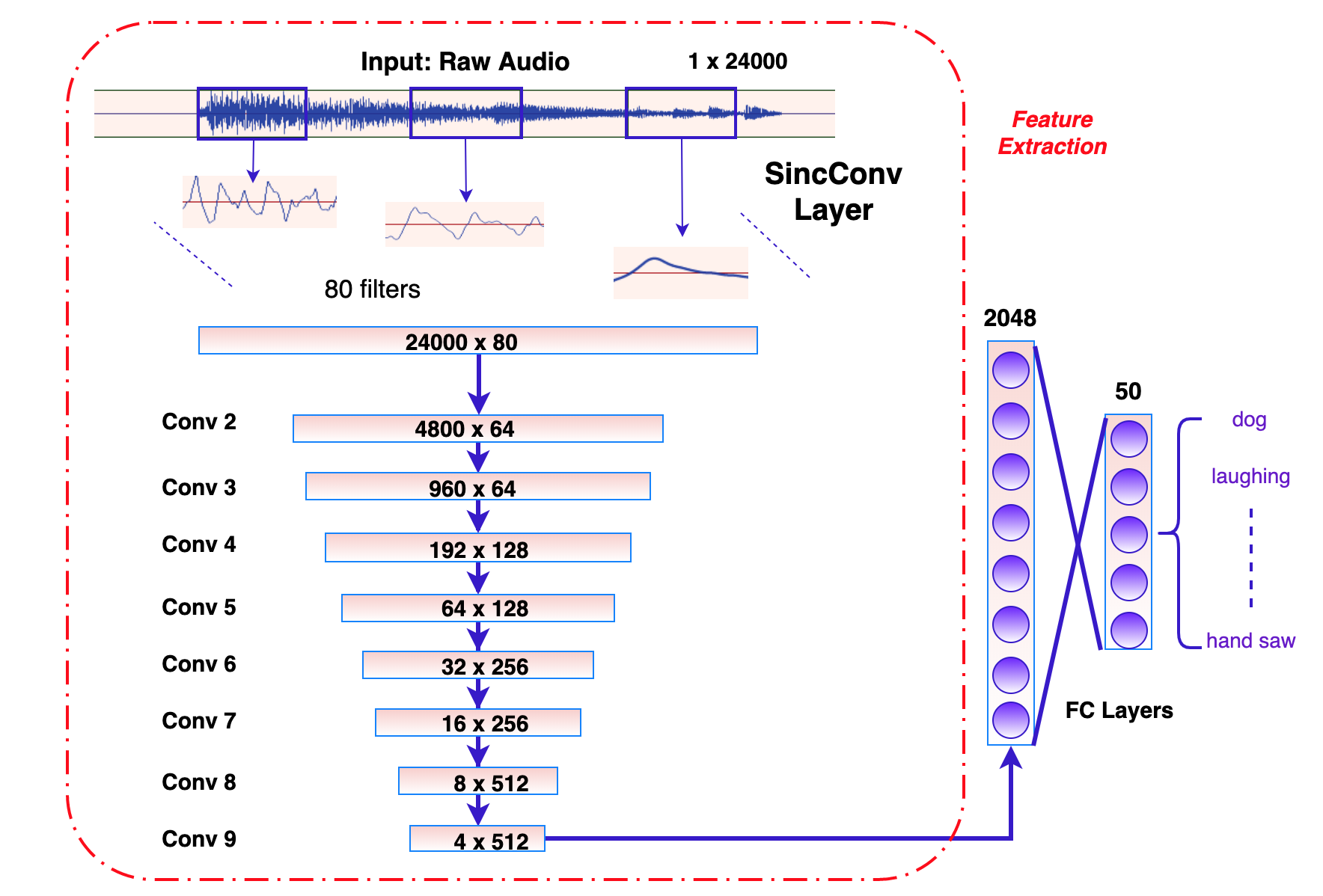


图14 基于SCNN网络的声学事件检测系统示意图(ESC50数据集下)

4.4 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。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。

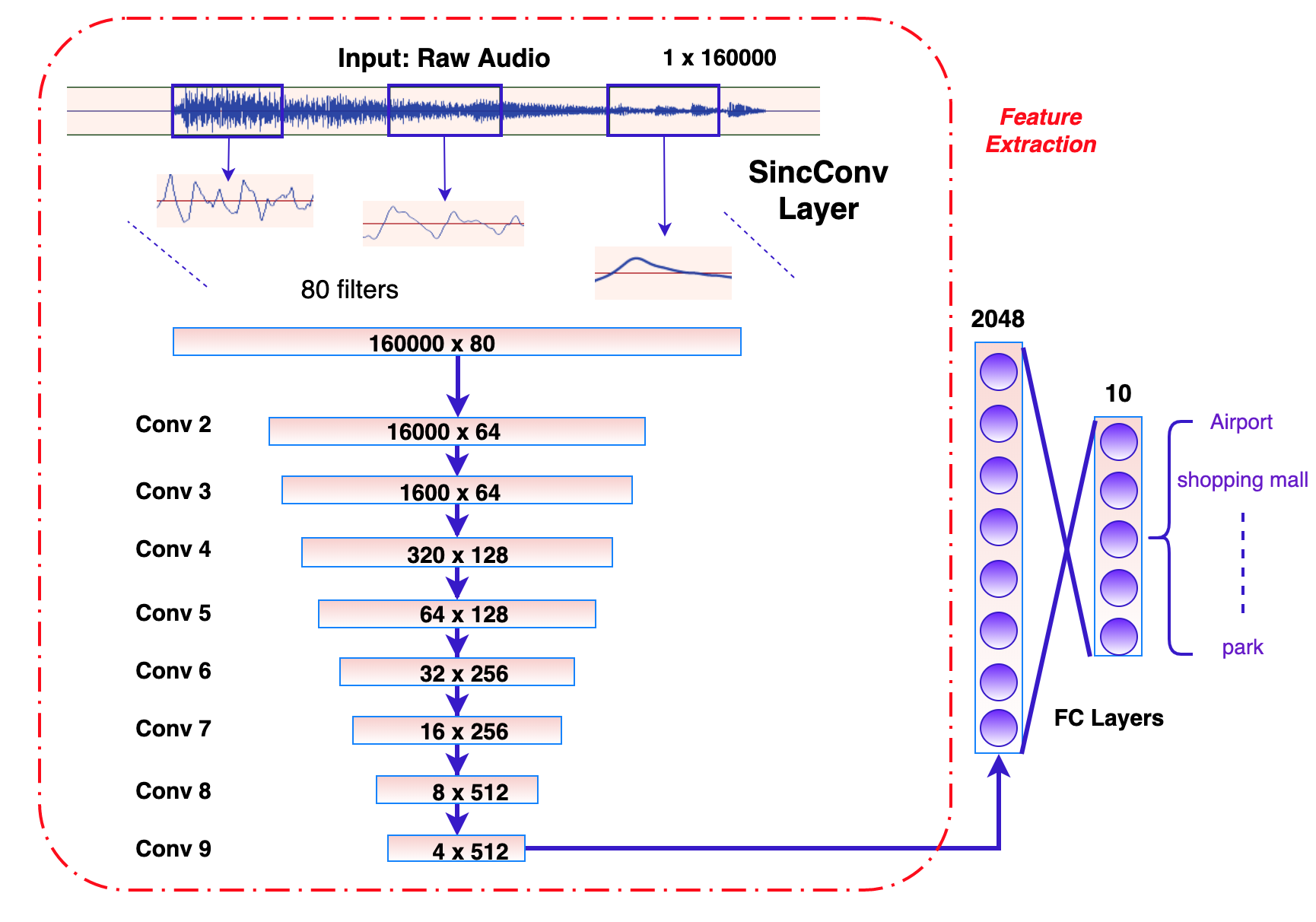


图15 基于SCNN网络的声学场景分类系统示意图(TUT Urban Acoustic Scenes 2018，Development dataset数据集下)

4.5 实验设置及结果分析对比

4.5.1 基于SCNN网络的声学事件检测系统实验设置

4.5.1.1 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。

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):

•动物的声音；

•自然的音景和水声；

•人（非语音）声音；

•室内/家庭声音；

•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。

4.5.1.2 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ESC50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。Batch size设置为64，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。

4.5.2 基于SCNN网络的声学场景分类系统实验设置

4.5.2.1 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。

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：

•机场-airport；

•室内购物中心-shopping mall；

•地铁站-metro station；

•步行街-street pedestrian；

•公共广场-public square；

•街道交通-street traffic；

•乘坐电车-tram；

•乘坐公共汽车-bus；

•乘坐地铁-metro；

•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。

4.5.2.2 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。

4.5.3 实验结果分析及对比

4.5.3.1 实验结果分析

表2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%)。

|  |  |
| --- | --- |
| 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2

表3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。

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

表3

图16和图17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18和图19分别为两个数据集下分类任务的混淆矩阵。

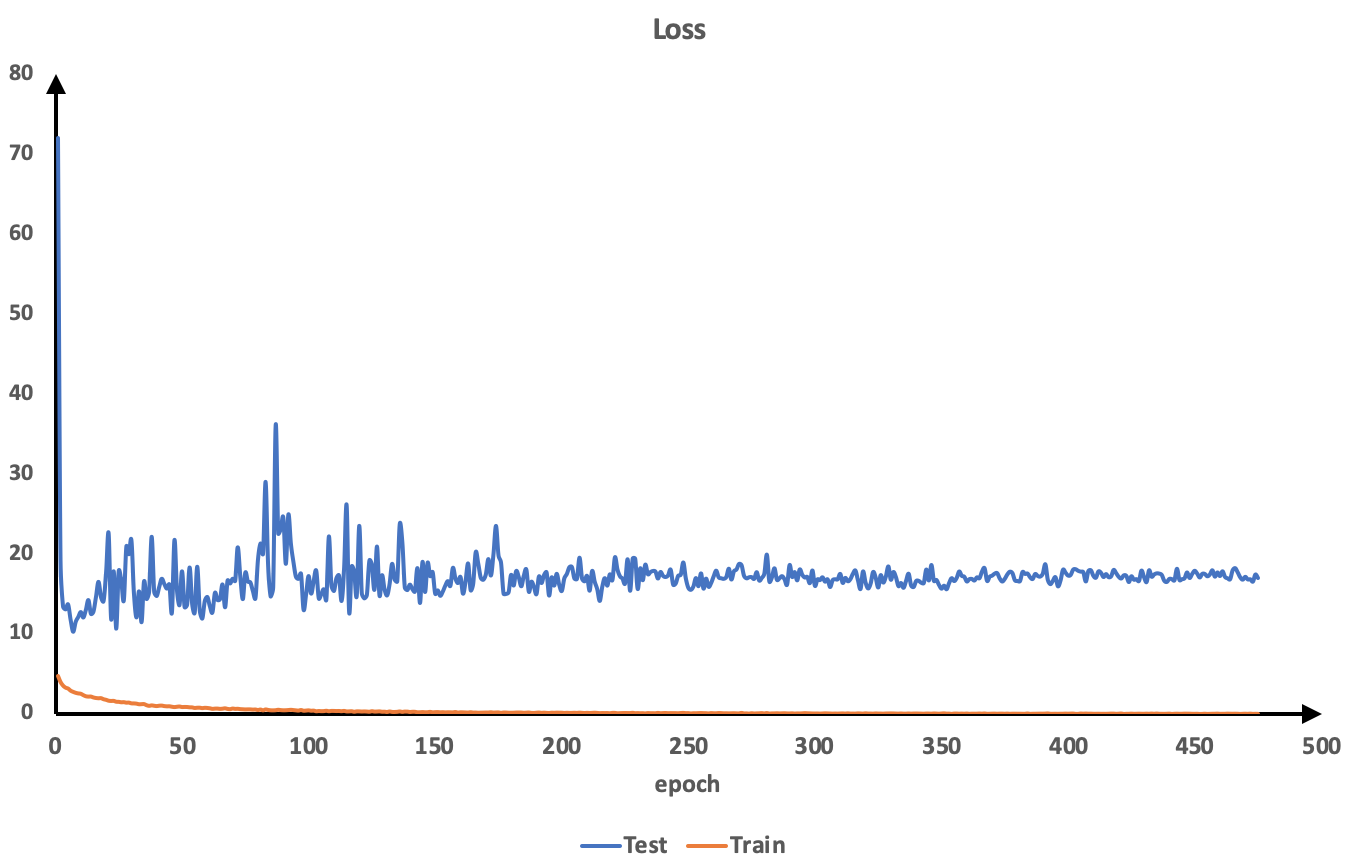


图16 训练阶段Loss情况

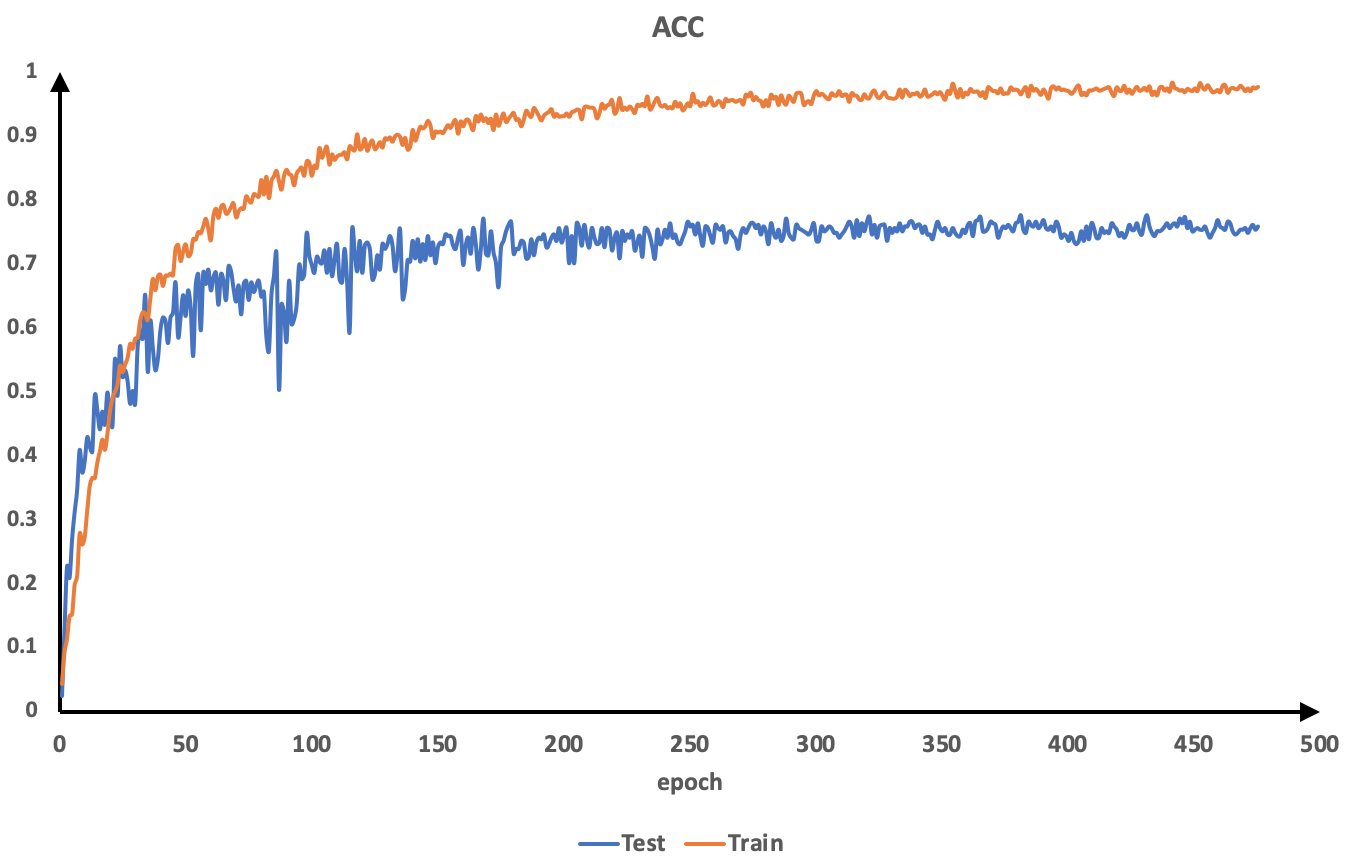


图17 训练阶段Accuracy情况

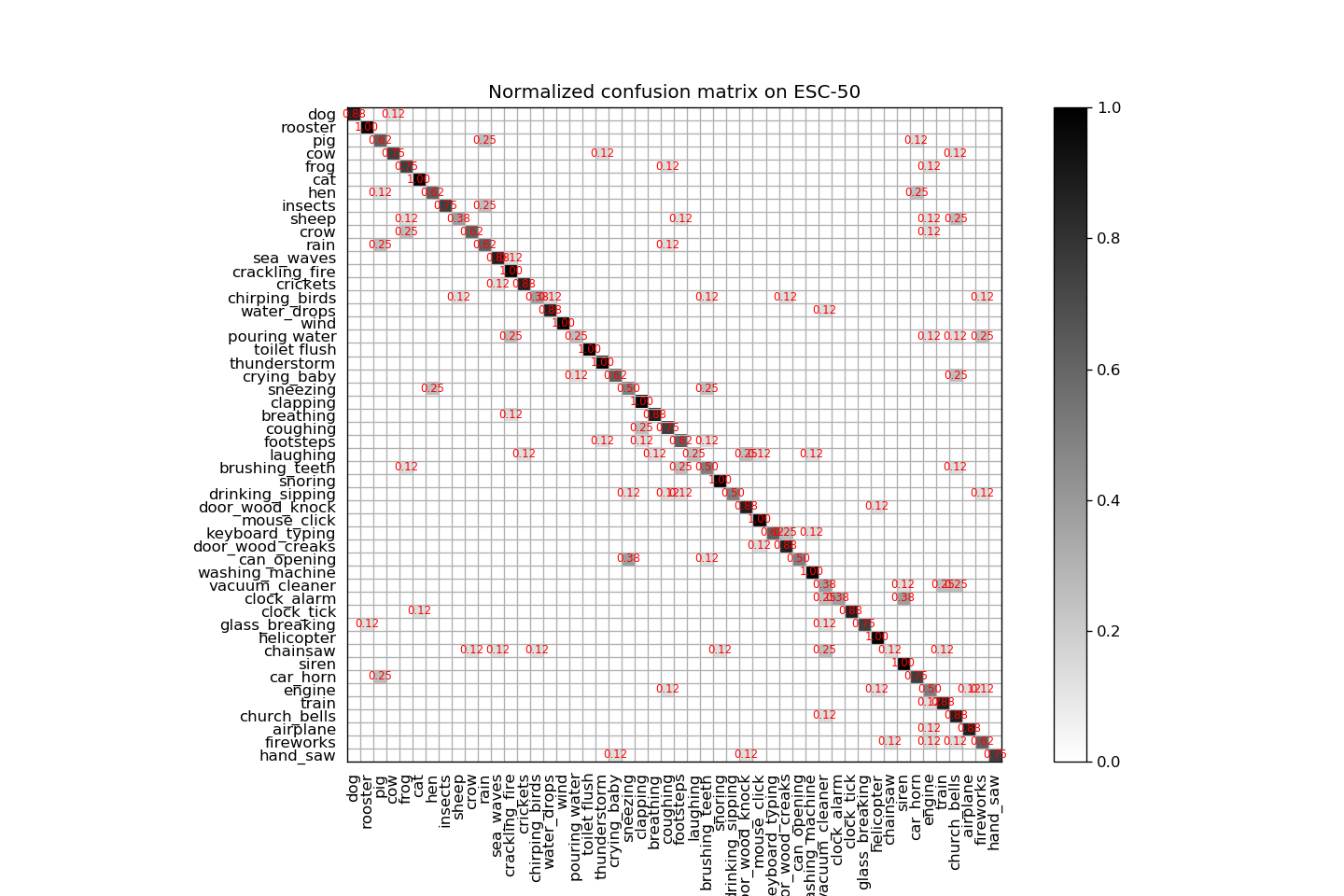


图18 ESC-50数据集上混淆矩阵

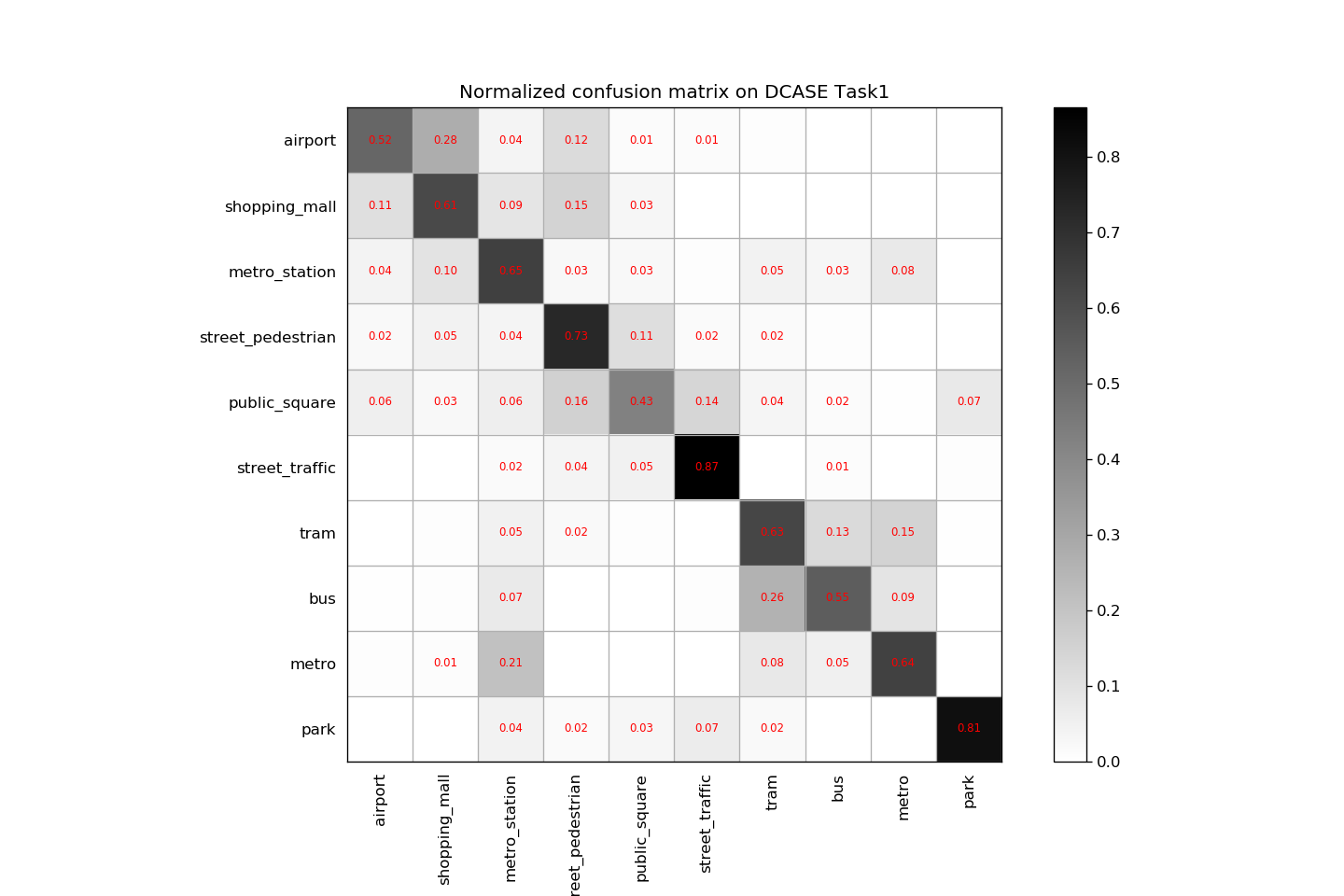


图19 DCASE Task1数据集上混淆矩阵

4.5.3.2 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16。累积频域响应图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16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17。

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18。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。

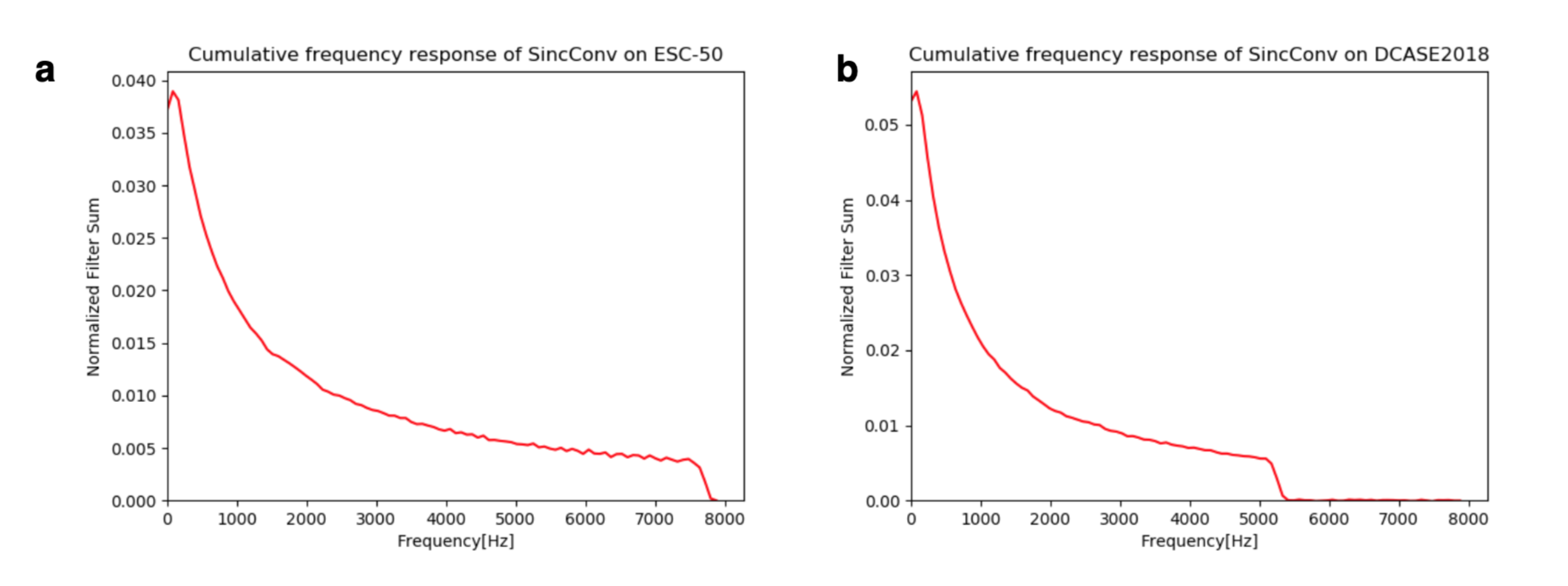


图16 SincConv卷积层在ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)下学习到的累积频域响应图。

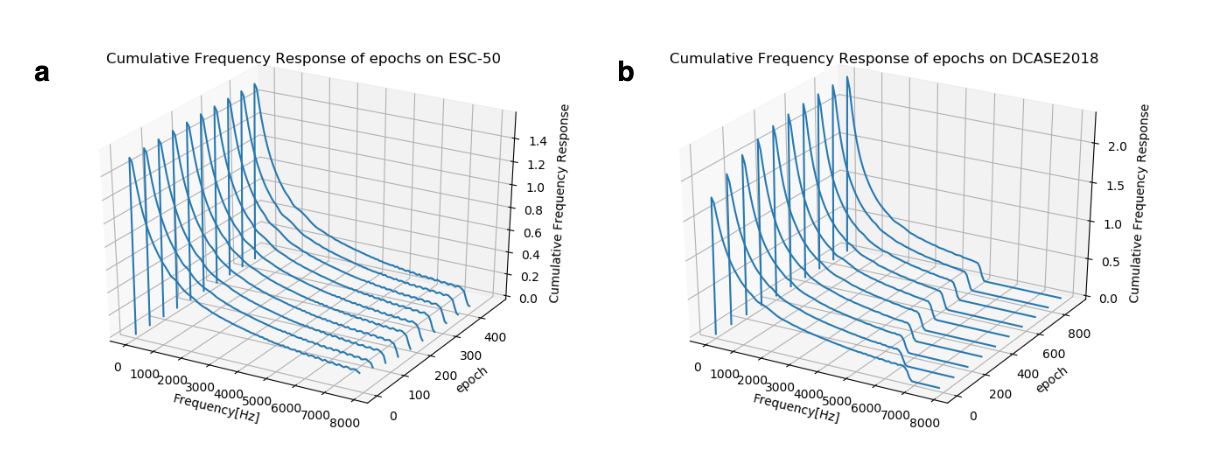


图17 随着训练的进行，在ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)下学习到累积频域响应图变化过程。x轴为频率(Hz)，y轴为Epoch(每50个显示)，z轴为频率响应。

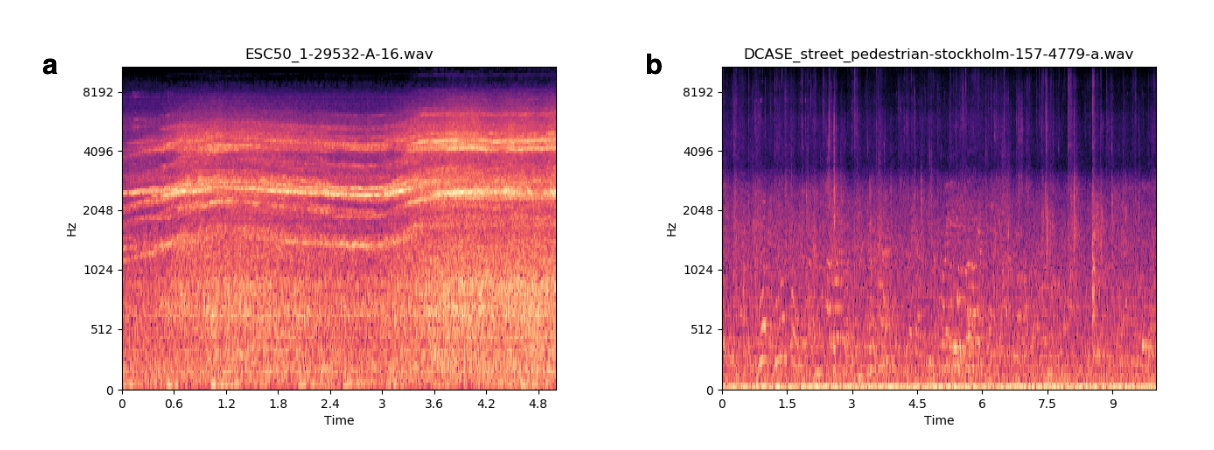


图18 ESC-50数据集(a)和TUT Urban Acoustic Scenes 2018, Development dataset数据集(b)中随机两组数据的对数梅尔频谱图

4.6 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。

第五章 基于SCNN网络的说话人识别系统

5.1 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节介绍相关实验及结果分析。

5.2 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，并且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，目标是找到听起来最接近于来自音频样本内的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，同时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。

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。前面讨论的变化和退化使这变得复杂，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。

5.3 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19。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)，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。

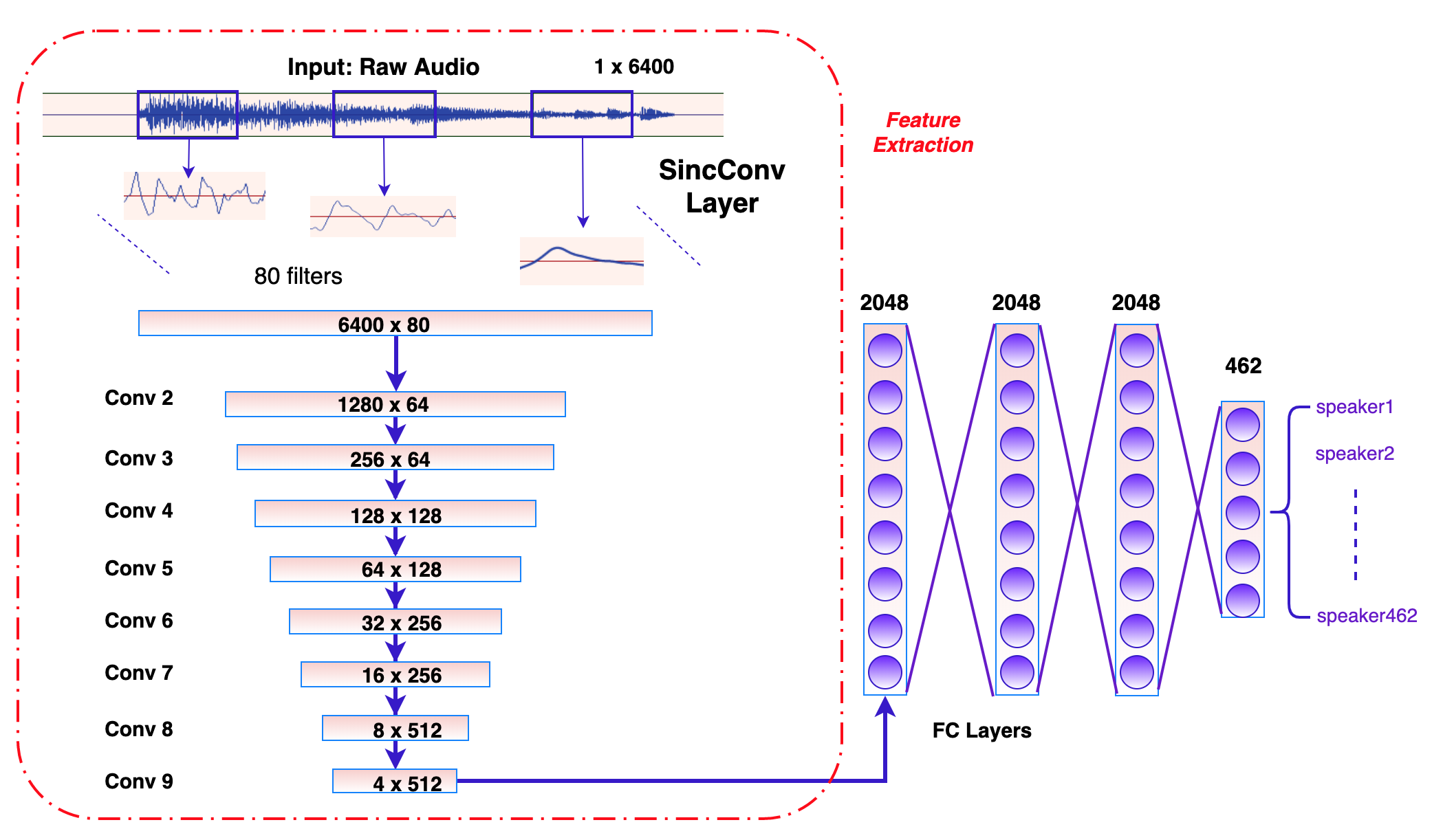


图19 基于SCNN网络的说话人识别系统示意图(TIMIT数据集下)

5.4 实验设置及结果分析对比

5.4.1 实验设置

5.4.1.1 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。

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。

5.4.1.2 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，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。

5.4.2 实验结果分析及对比

表4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。

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

表4

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20。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。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。

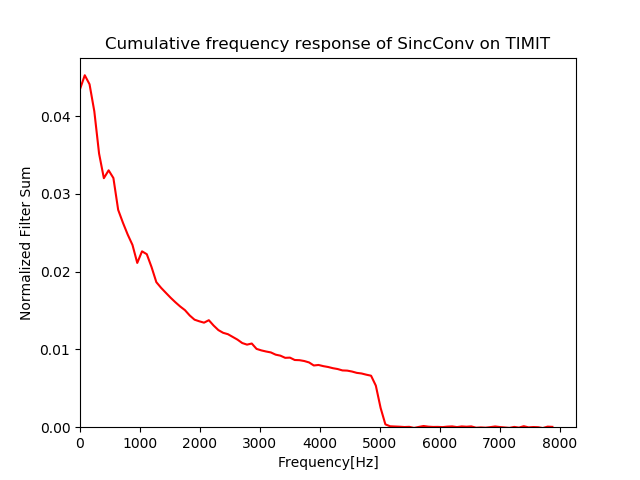


图20 SincConv卷积层在TIMIT数据集下学习到的累积频域响应图。

5.5 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。

第六章 总结与展望

6.1 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。

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。

6.2 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。

参考文献

致谢