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1章 引言

1.1研究背景

1.1.1机器听觉

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。

1.1.2声学事件和声学场景分析

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，鸟鸣或门铃等。声学事件只有一个来源，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-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。

1.1.3说话人识别

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。

1.1.4声学特征提取

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，我们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。

1.1.5深度神经网络

深度学习发展至今，经历三次发展浪潮：20世纪40年代到60年代深度学习的雏形出现在控制论（cybernetics）中，20世纪80年代到90年代深度学习变现为联接主义（connectionism）。直到2006年，才真正以深度学习之名复兴。

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。

1.2研究现状

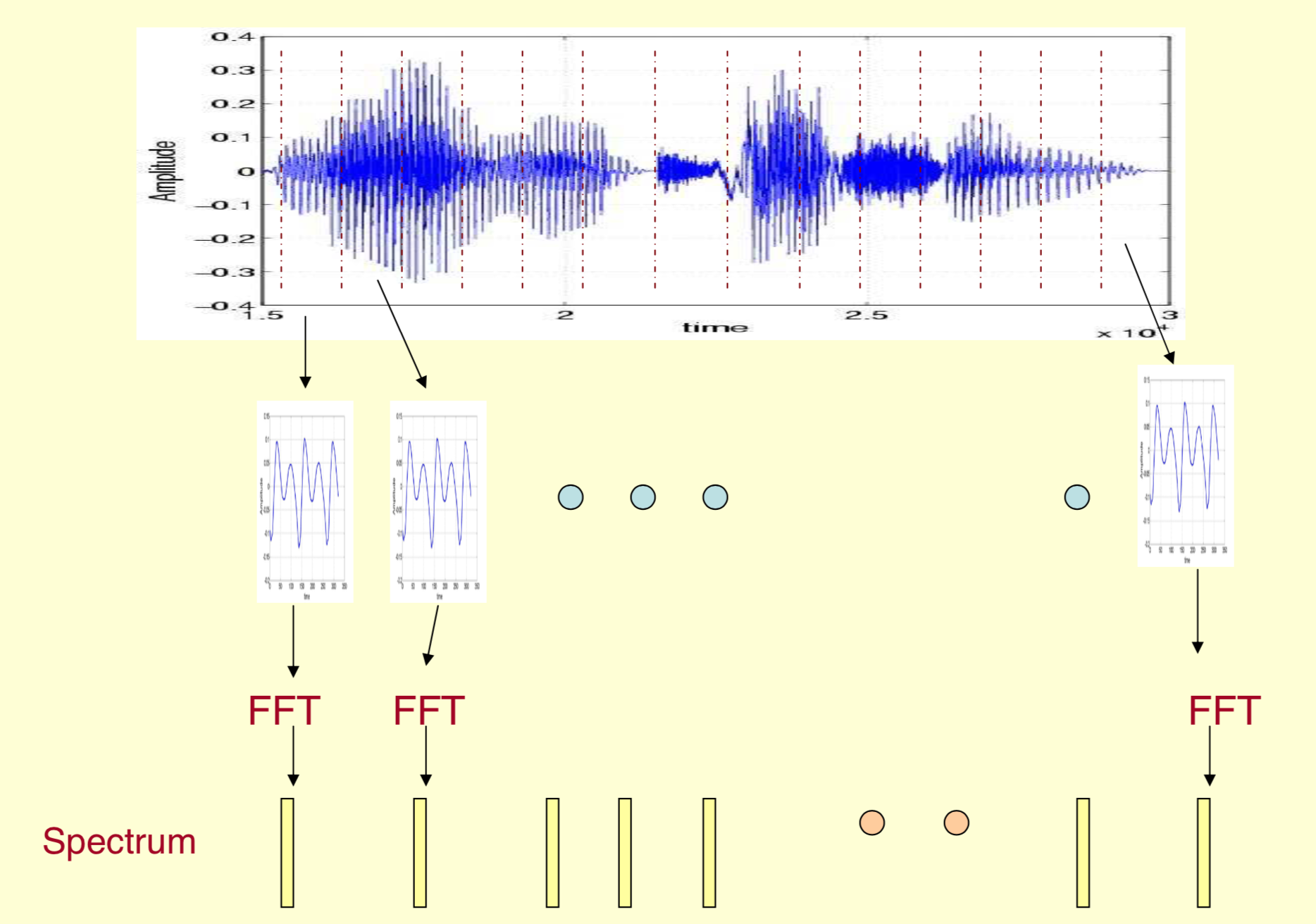
1.2.1人工声学特征提取方法概述

至今，许多音频处理任务广泛采用人工声学特征提取方法。经典的方法有声谱图(Spectrogram)、Log-Mel能量、梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)等。这些人工方法利用了信号处理、声学等知识进行设计，对于不同的音频处理任务有不同的人工特征提取手段。

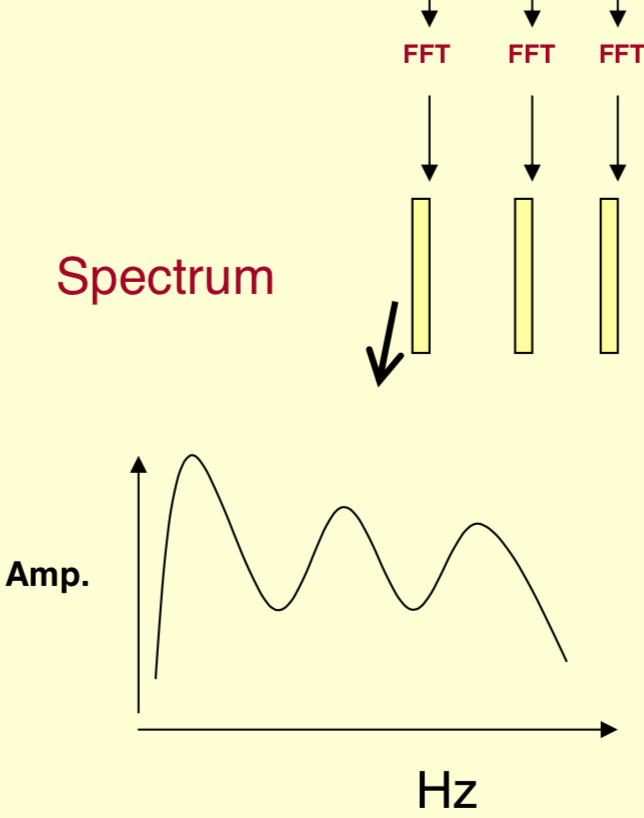
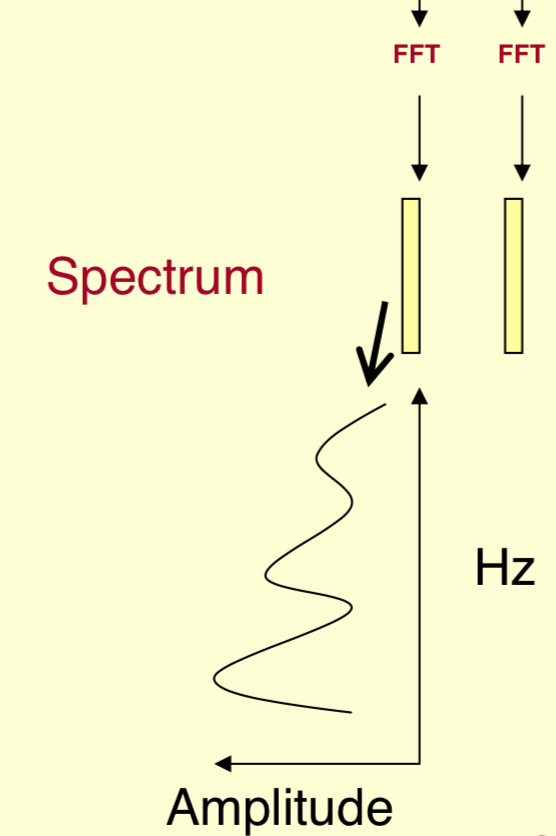
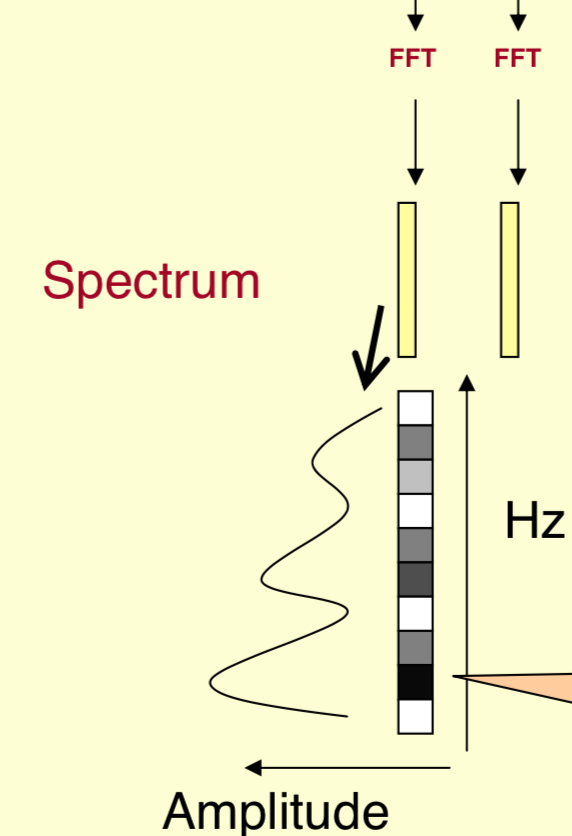
1.2.1.1声谱图

声谱图，即声音频谱图，是通过处理时域信号得到的频谱图，可以用来反映音频的时域和频域特性。对于语音的研究可以称为语谱图，即语音频谱图。

声谱图的横坐标是时间，纵坐标是频率，坐标点值为声音数据能量。在信号领域，傅里叶变换(FFT)常常是联系时域和频域的纽带，在频域里可以更方便的进行一些分析。傅里叶变换主要针对的是平稳信号的频率特性分析，即具有一定周期性的信号，因为傅里叶变换采取的是有限取样的方式，所以对于取样长度和取样对象有着一定的要求。在斯坦福大学的课程中，声谱图的具体提取方法介绍如下。



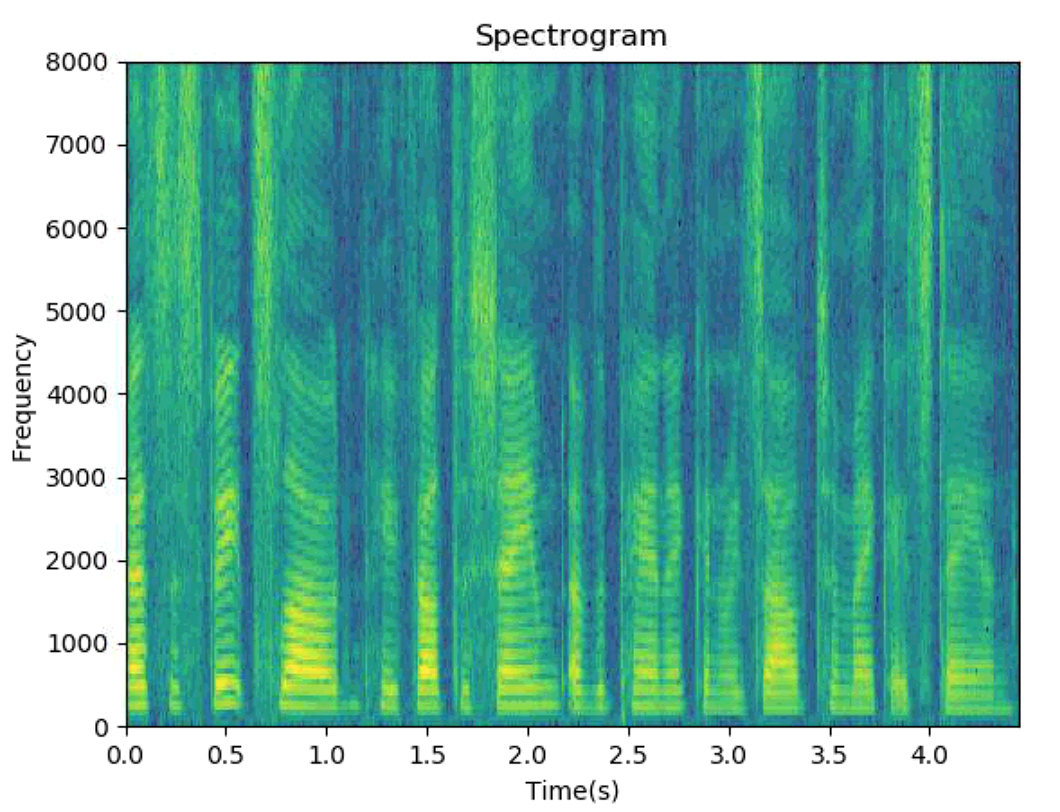
首先，对音频进行加窗分帧处理，常用的窗函数有矩形窗、三角窗、汉宁窗、海明窗以及高斯窗等，这里汉宁窗和海明窗使用较多。每帧音频通过短时傅立叶变换计算，对应一个个频谱，频谱用来表示频率与能量的关系。

我们将其中一帧音频的频谱通过坐标表示出来，横坐标是频率，纵坐标是能量，如上左图。然后，将左图的频谱旋转90度，得到中间的图。最后把这些幅度映射到一个灰度级表示（也可以理解为将连续的幅度量化为256个量化值），0表示黑，255表示白色。幅度值越大，相应的区域越黑。这样就得到了最右边的图。该操作是为了增加时间这个维度，这样就可以显示一段音频而不是一帧语音的频谱，而且可以直观的看到静态和动态的信息。



这样就得到了如上的随时间变化的频谱图，这就是描述音频信号的spectrogram声谱图。由于是采用二维平面来表达三维信息，所以能量值的大小一般是通过不同的颜色来表示的，而较少采用灰度图，这样视觉效果更加明显。如下图，颜色越鲜艳，表示该点的声音能量越强。



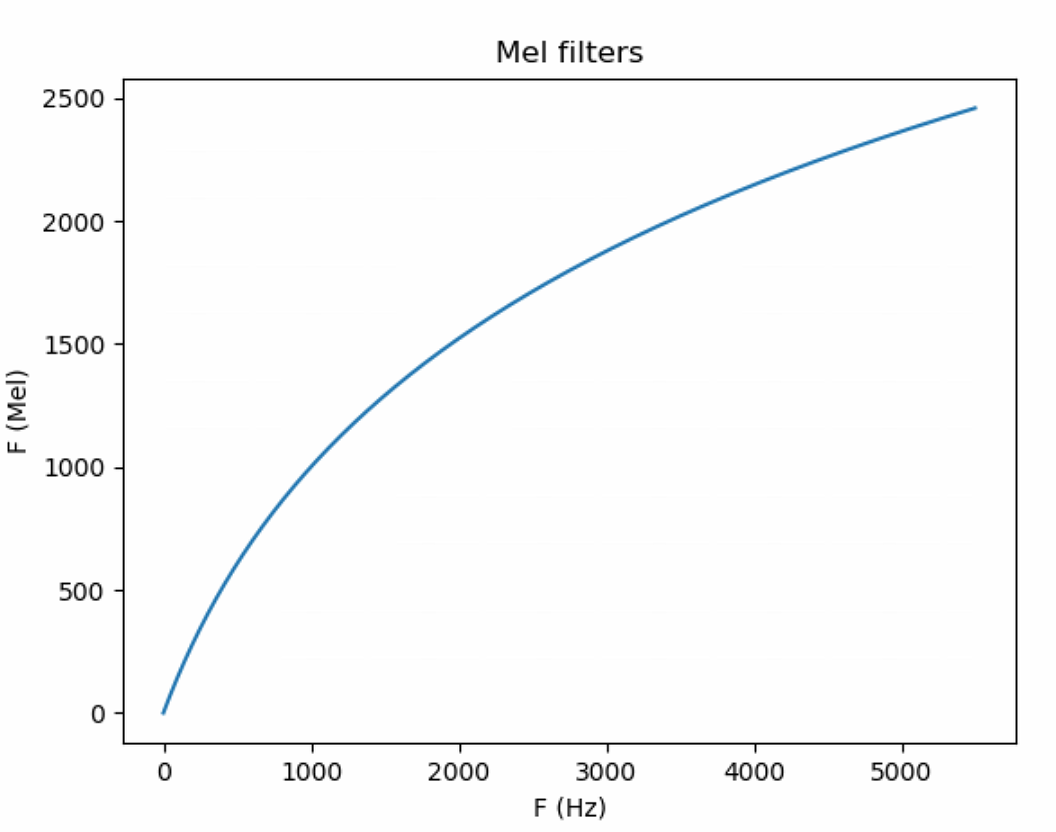
1.2.1.2 Log-Mel能量

上面介绍了声音频谱图的提取方法。在实际使用时，频谱图有三种，即线性振幅谱、对数振幅谱和自功率谱。线性振幅谱的纵坐标有明确的物理量纲，声谱图属于线性振幅谱。而对数振幅谱中各谱线的振幅都作了对数计算，其纵坐标的单位是dB（分贝）。这个变换的目的是使那些振幅较低的成分相对高振幅成分得以拉高，以便观察掩盖在低幅噪声中的周期信号。

对于人类听觉感知的实验表明，人类听觉的感知只聚焦在某些特定的区域，而不是整个频谱范围。Mel频率分析就是基于人类听觉感知实验的。实验观测发现人耳就像一个滤波器组一样，它只关注某些特定的频率分量（人的听觉对频率是有选择性的）。但是这些滤波器在频率坐标轴上却不是统一分布的，在低频区域有很多的滤波器，他们分布比较密集，但在高频区域，滤波器的数目就变得比较少，分布很稀疏。

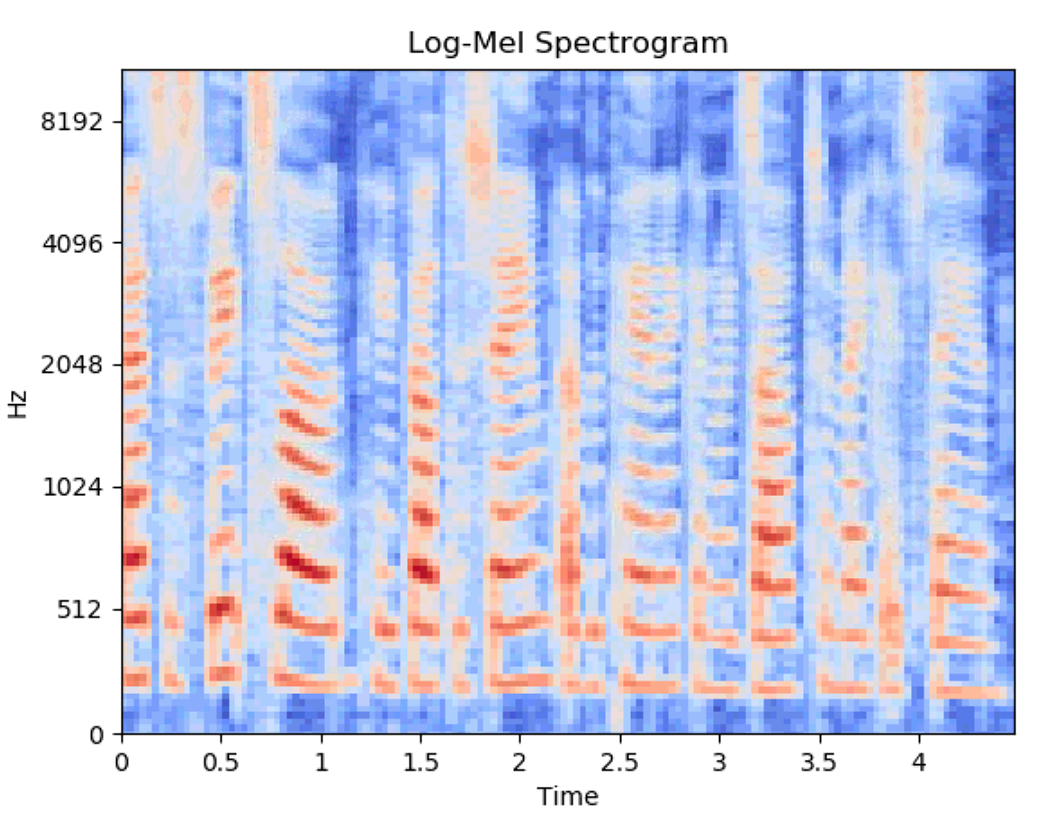
按照下面的公式可将普通频率转化为Mel频率：

从下图可以看到，它可以将不统一的频率转化为统一的频率，也就是统一的滤波器组。

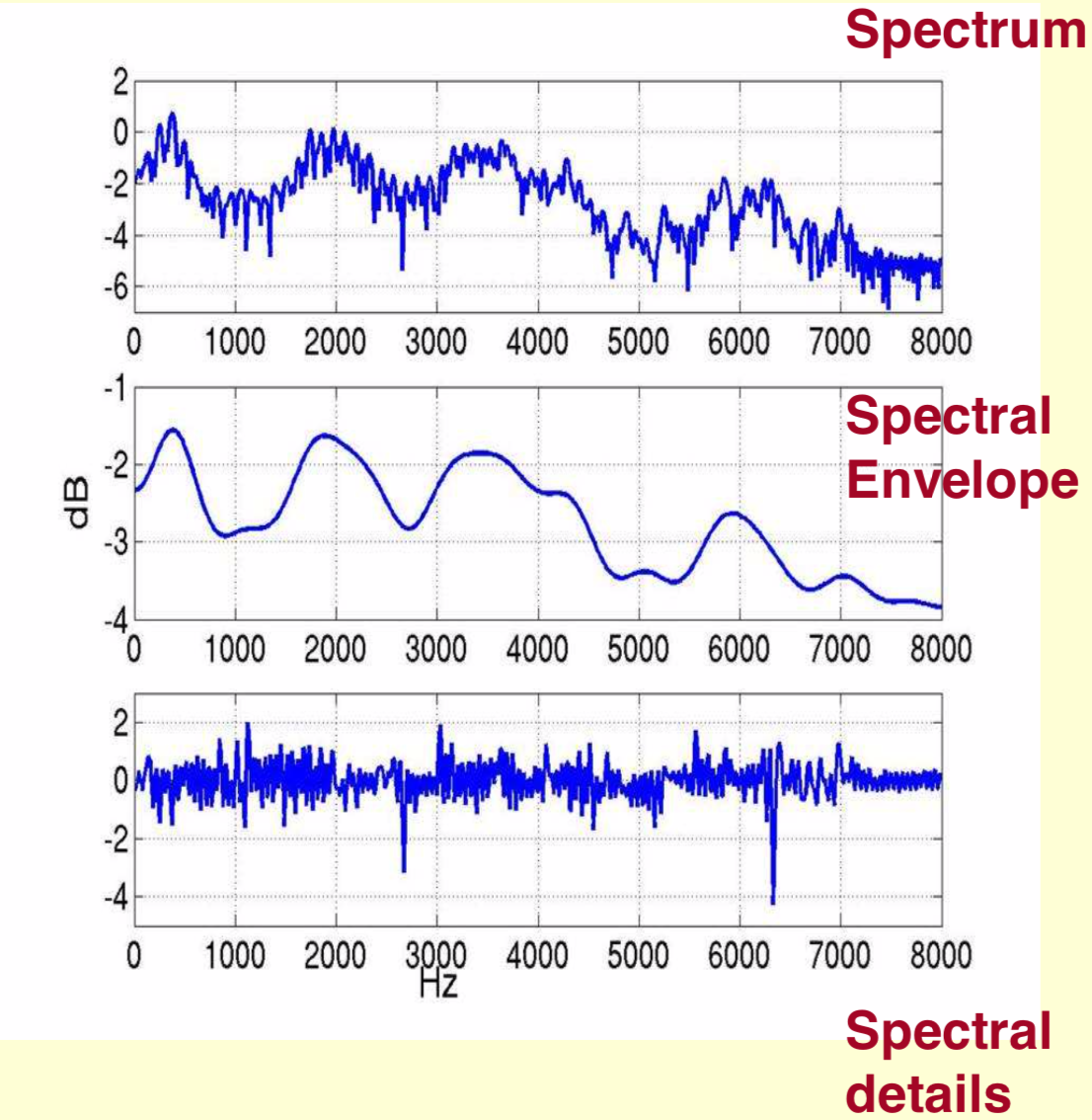


在Mel频域内，人对音调的感知度为线性关系。举例来说，如果两段音频的Mel频率相差两倍，则人耳听起来两者的音调也相差两倍。

Log-Mel能量，也称为梅尔对数能量谱，也以此为基础设计。首先对音频进行加窗分帧，然后对每一个短时分析窗，通过短时傅里叶变换得到对应的频谱，再通过Mel滤波器组得到Mel频谱，即可得到Log-Mel能量，如下图。



1.2.1.3 MFCC



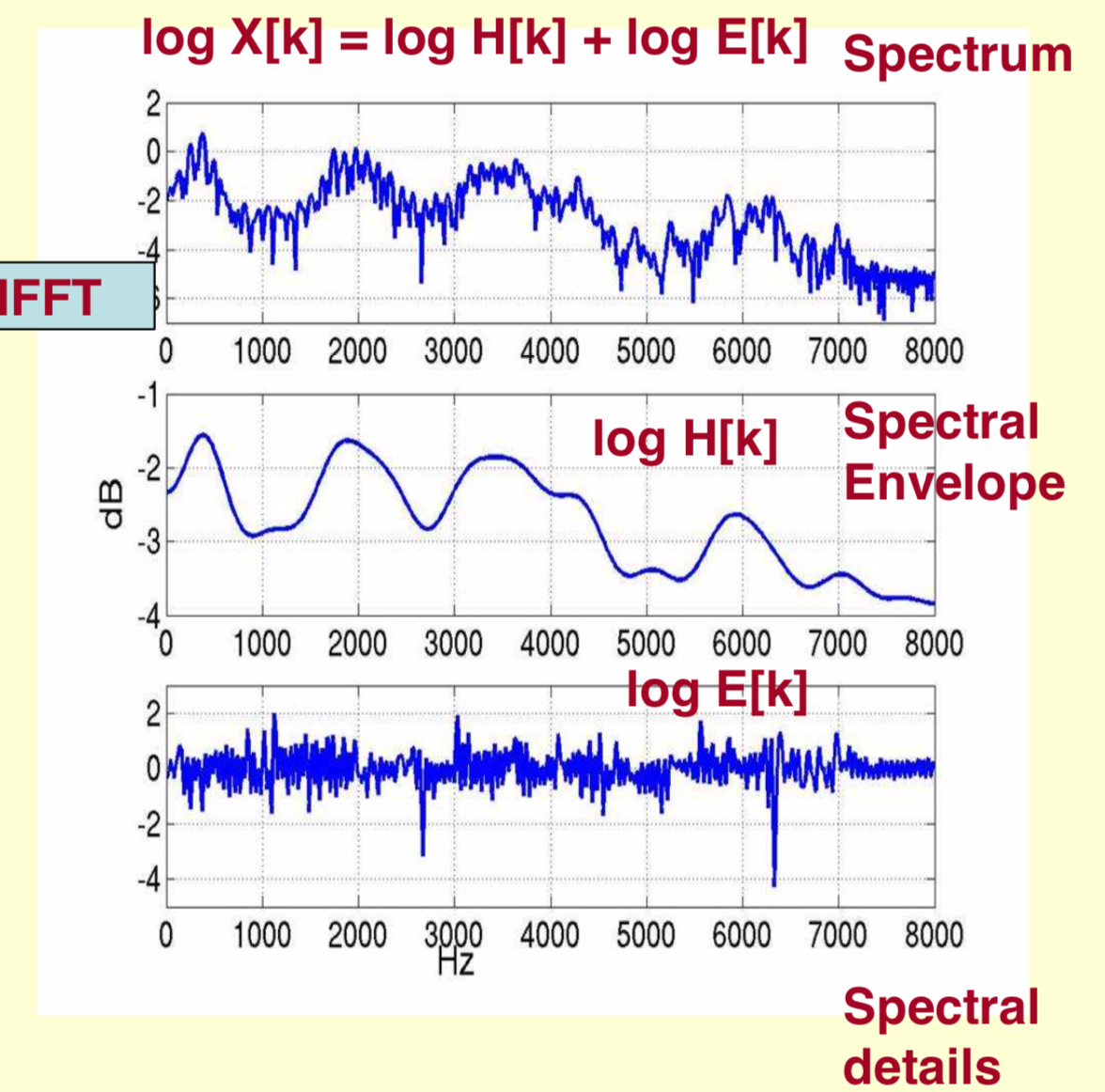
上图Spectrum为频谱图，我们可以通过1.2.1.1的方法得到。它可以分成频谱的包络(Spectral Envelope)和频谱的细节(Spectral details)两个部分。这里采用的是对数频谱，单位是dB。我们经常需要得到频谱的包络来进行有意义的分析，这就需要将二者分离开。

我们可以把原始音频信号通过傅立叶变换得到的频谱看成：

如果只考虑幅度就是：

在两边取对数：

再取逆傅立叶变换得到：



我们已经知道，所以可以得到。那么由上图可以知道，是的低频部分，所以我们将通过一个低通滤波器就可以得到，也就是得到了频谱的包络。

以上分析方法叫做倒谱分析(Cepstrum)。如果在Mel频谱上进行倒谱分析（取对数，做逆变换，实际逆变换一般是通过DCT离散余弦变换来实现，取DCT后的第2个到第13个系数作为MFCC系数），获得Mel频率倒谱系数（MFCC），这个MFCC就是这帧音频的特征。MFCC被广泛应用在语音处理上。

1.2.2 基于神经网络的深度声学特征提取概述

近年来，随着神经网络在计算机视觉、语音识别等领域的成功，人们也越来越关注其在特征学习上的应用。通过向神经网络提供手工特征以提取其深度特征，或者直接用神经网络对原始音频进行特征提取，已经成为了一项重要的研究工作，下面简要介绍采用CNN和DNN的方法进行深度声学特征提取。

1.2.2.1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。

1.2.2.2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。

1.3 毕业论文整体执行完成情况概述

1.3.1 知识学习阶段

在知识学习和准备阶段，我针对机器听觉相关知识进行了学习，了解其背后的信号处理知识，学习Pytorch框架下的深度神经网络，Linux系统下的基本操作以及声学特征提取的各种方法等。

1.3.2 方案设计阶段

通过对比几种主流的方法，确定本次毕业论文的方案。

方案一：根据语谱图提取深度特征

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；

优点：思路简单，容易实现；

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。

方案二： 利用MFCC等特征提取深度特征

描述：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；

优点：较可靠，较容易实现；

缺点：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。

方案三： 利用原始信号提取深度特征

描述：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：可以保留完整的音频信息，可靠，属于主流设计方案;

缺点：网络设计较难。

考虑到上述方案的优缺点，拟采用方案三进行本次毕业论文设计，即直接向深度神经网络馈送原始音频信号，这种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。

1.3.3 实验阶段

在实验阶段，针对所设计的方案进行网络结构搭建，数据集整理，代码编写，性能测试等工作。

1.3.4 系统调试、优化及性能分析阶段

根据实验测试结果优化网络结构，改善方案细节，对于所测试的系统性能进行分析。

1.3.5 毕业论文工作总结阶段

总结实验结果和毕业论文相关工作。

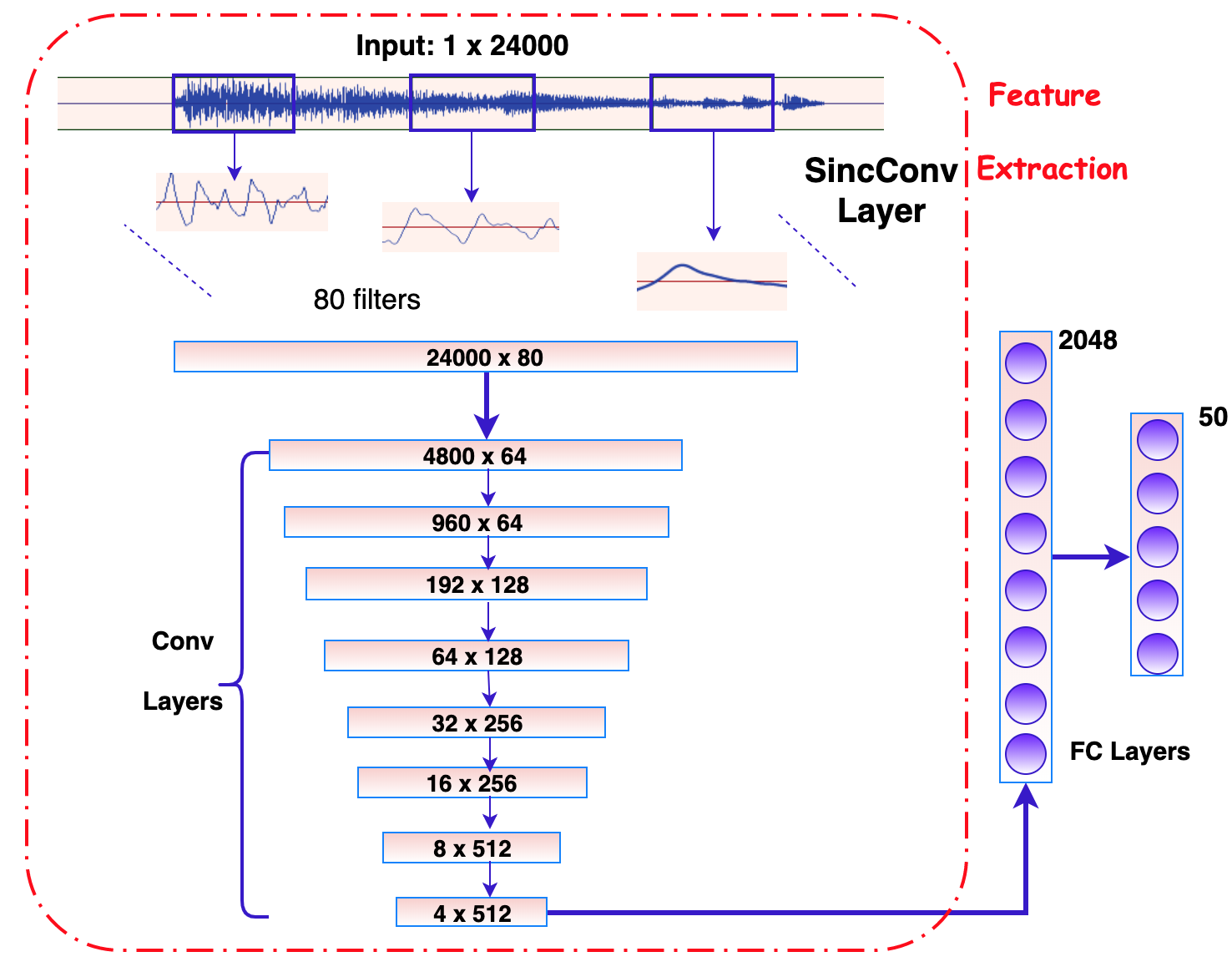
1.4 毕业论文组织结构安排

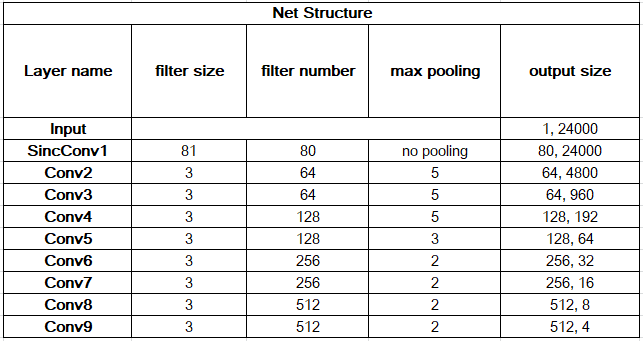
第2章 研究方案

2.1设计思路

本次毕业论文训练采用基于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VGGNet，第一层采用SincConv取代标准卷积层结构，下图1是基于声学事件检测任务下的网络结构图。层之间采用Batch Normalization进行批标准化，激活函数采用leaky relu函数，如图2。整体模型我称之为SCNN(Convolution Neural Network with SincConv Layer)。

图3是每层网络的具体参数情况。





2.2 SincConv卷积层

上图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卷积层提出一种有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：

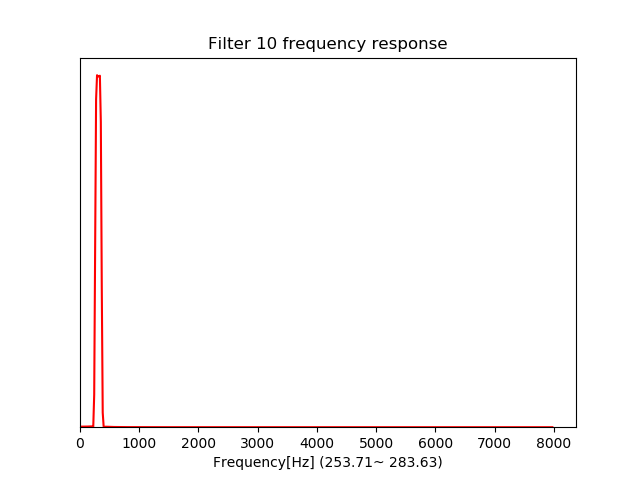
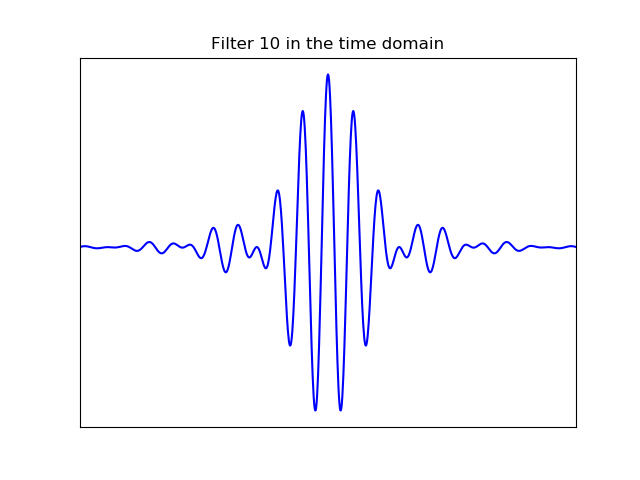
截止频率可以在范围内随机初始化，其中代表输入信号的采样频率。一种比较好的初始化方式，滤波器可以采用梅尔滤波器组的截止频率初始化，这种设计更加考虑到人的听觉特点，在低频区域设置了更多滤波器。这里的不需要限制在奈奎斯特频率以下，因为在训练的过程中和作为参数可以自动满足该条件。而且，该层网络不需要学习每个滤波器的增益，这些参数可以在后续网络层学习得到。

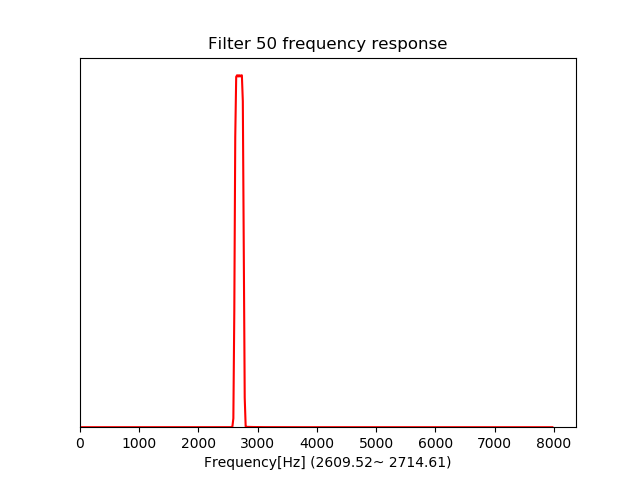
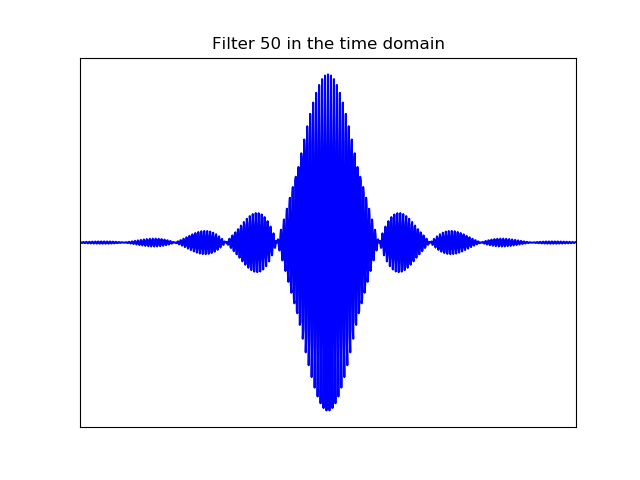
理想的带通滤波器（即通带完全平坦并且阻带衰减无限的滤波器）需要参数的个数无限多。因此，任何函数都只能实现理想带通滤波器的近似，因为其在通带处存在波动，同时阻带的衰减有限。为了减轻这种频谱能量泄漏的问题，通常会采用加窗的方法，通过将截断函数与窗口函数相乘来实现窗口化，窗口函数可以平滑边缘的突然不连续性：

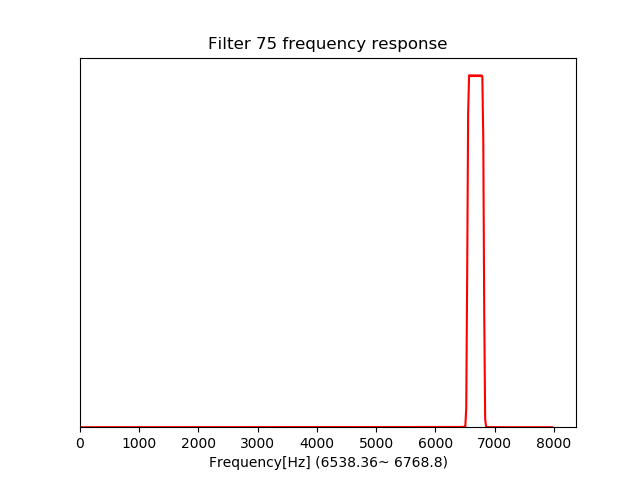
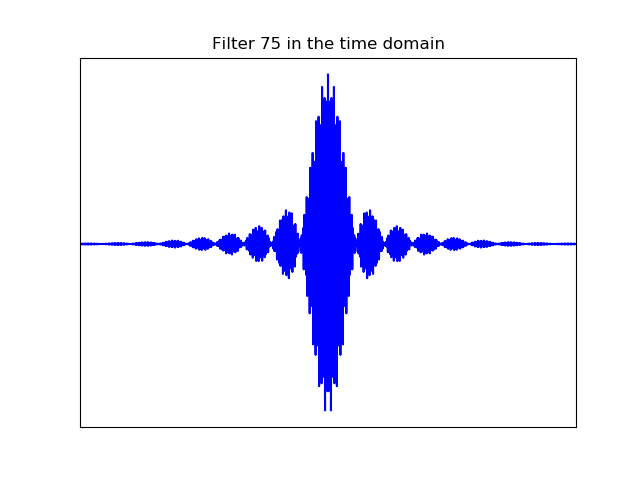
SincConv卷积层中使用的是Hamming窗，定义如下：

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。

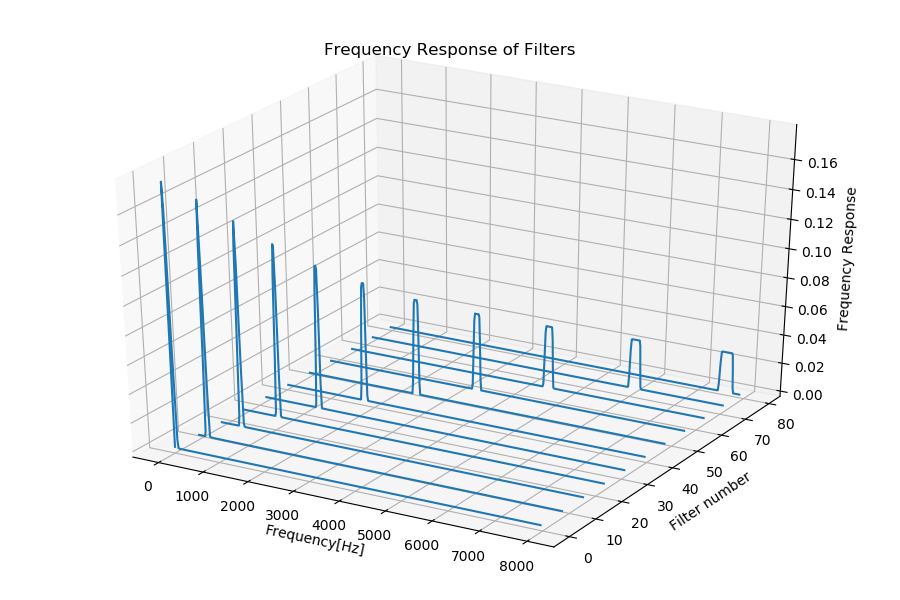
第一层SincConv卷积可以看作将原始音频信号通过若干个带通滤波器，输出的是不同频段的时域信息，供后段网络进一步进行特征提取。如图为几个学习到的滤波器的时域和频域特性图。



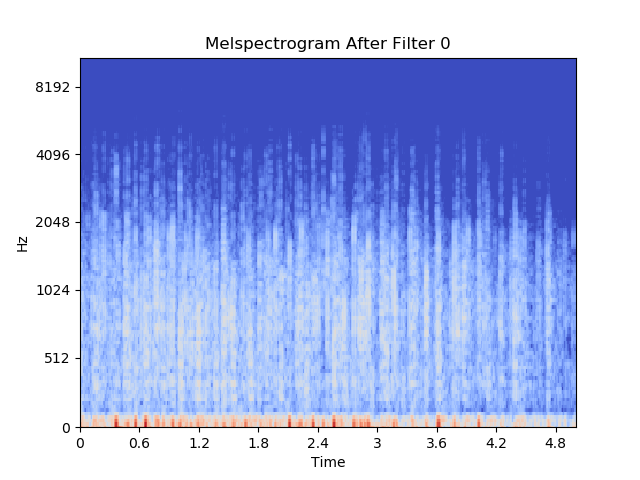
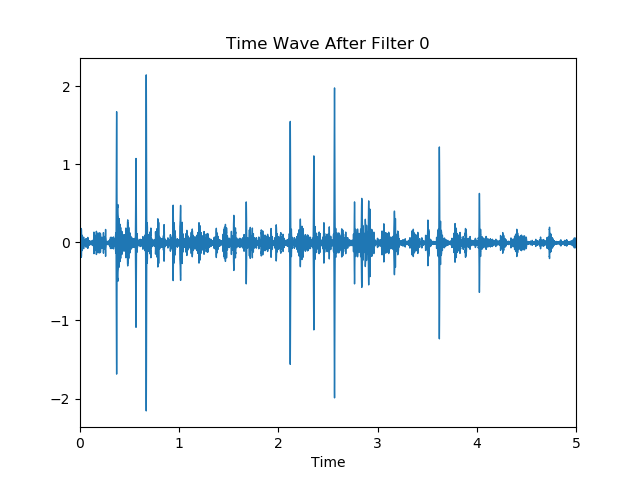


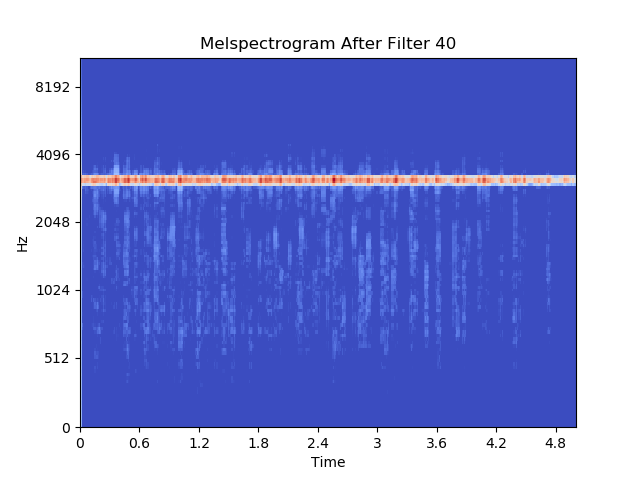
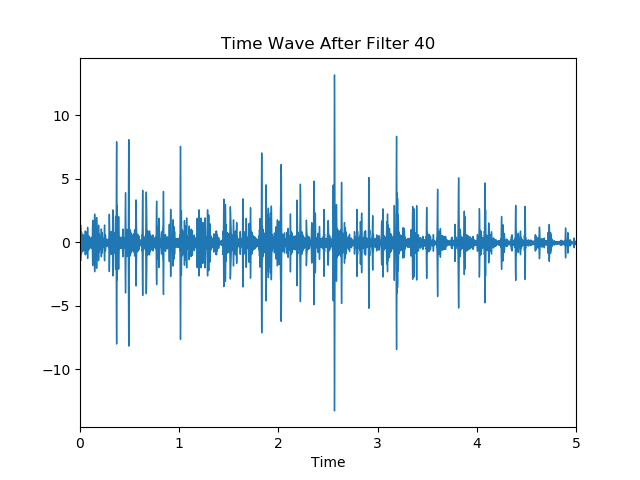


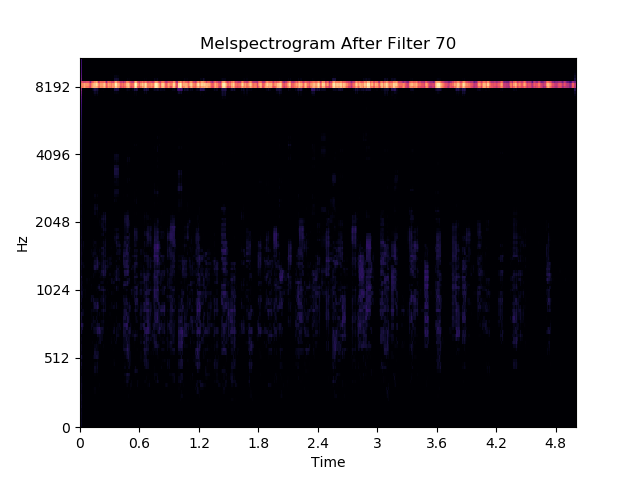
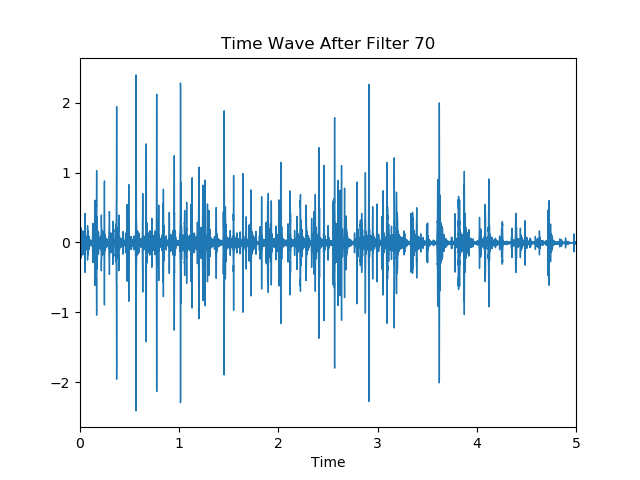
下图画出所有的滤波器的频域响应图，可以直观地看到学习到的不同滤波器的频段响应。



在一段音频数据上，通过上述滤波器组可以得到不同的音频信息，通过绘制时域波形和log-mel频谱图来观察其特性。







2.3后端网络设计

如图，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(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，增加网络的泛化能力。以输入音频信号为例(采样频率为16kHz，时长1.5s)，通过该特征提取网络输出的深度特征维数为。

2.4 相关方法概述

2.4.1 激活函数

2.4.2 对于梯度问题的处理

2.4.3 损失函数

2.4.4 优化器算法

2.5 本章小结

第3章 实验及相关配置

3.1 实验数据集

为了测试所设计的网络的性能，将其分别于声学事件检测任务和说话人任务上进行测试。以下介绍用到的数据库。

3.1.1 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。

3.1.2 TUT Urban Acoustic Scenes 2018 dataset

TUT Urban Acoustic Scenes 2018 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。

3.1.3 Urbansound8K 数据集

Urbansound8K 数据集是Urbansound中用于城市声音分类的数据集，包含8732个标记的声音片段(每个片段时长在4s以内)，共分为10类声音:

•空调-air conditioner；

•汽车喇叭-car horn；

•儿童游戏-children playing；

•犬吠-dog bark；

•钻孔-drilling；

•发动机怠速-engine idling；

•枪击-gun shot；

•手提钻-jackhammer；

•警报器-siren；

•街头音乐-street music。

Urbansound8K 数据集被预分为10折，便于交叉验证。

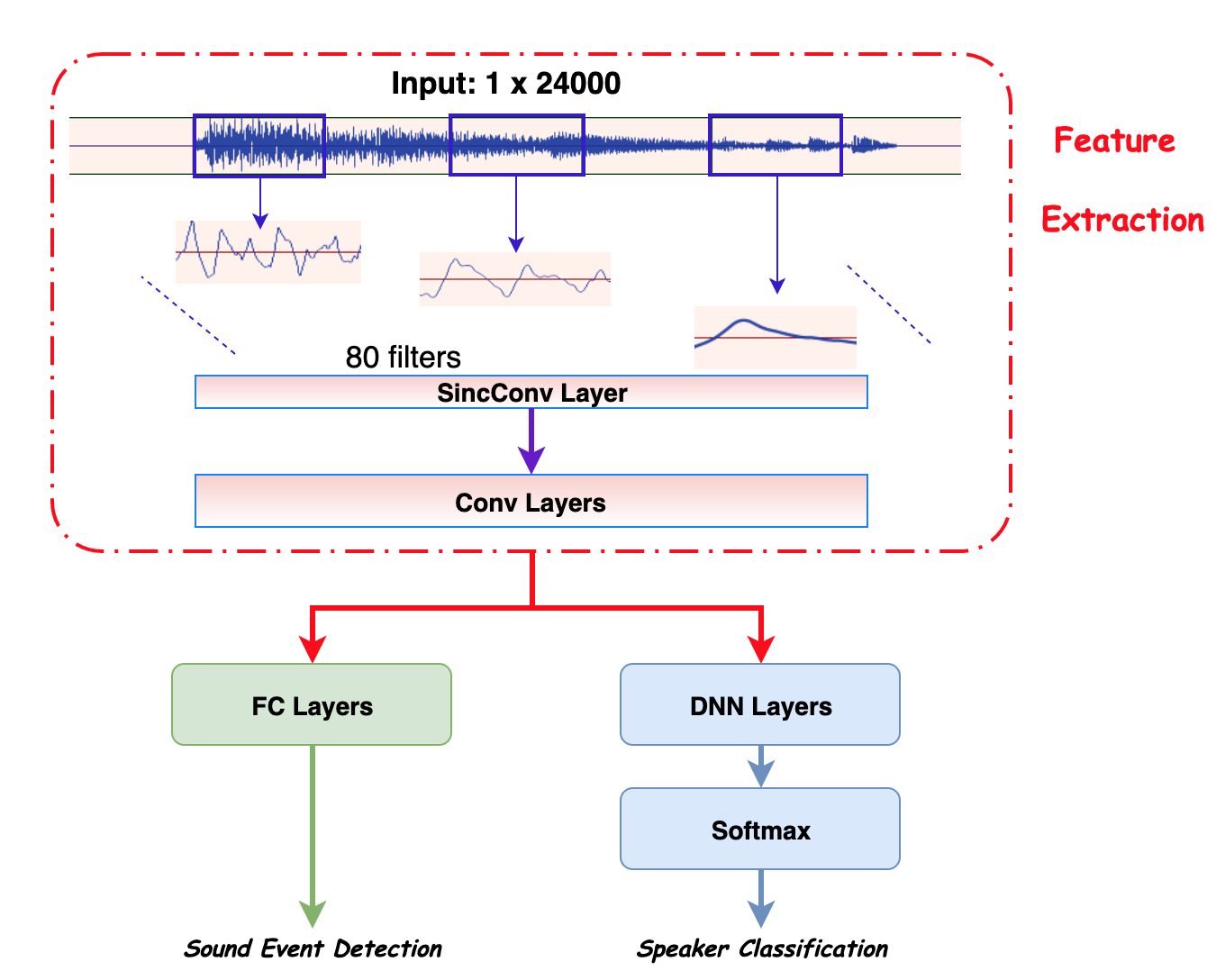
3.1.4 TIMIT 数据集

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。

3.2 实验运行环境

3.3 实验具体方法及参数设置

在3.1所述数据集中进行性能测试，分别基于声学事件检测任务和说话人识别任务，所采用的网络结构如下图。



前端特征提取部分采用2.1所述网络结构，区别是在声学事件检测任务上，特征提取操作之后紧接着是全联接层进行分类；在说话人识别任务中，特征提取操作之后连接的是DNN网络和Softmax层进行分类。

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Mel bank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，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。

对于声学事件检测任务，采用随机选取原始音频中1.5s的连续信息作为输入，在测试阶段，我采用了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。  
 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分类器获得帧级别分类和音素级别分类。使用RMSprop优化器进行训练，学习率为，，以及batch size设置为128。

3.4 本章总结

第4章 实验结果分析

4.1 实验过程分析

4.2 模型性能测试

4.2.1 模型性能评估指标

4.2.2 模型性能

首先，我将该特征提取模型于ESC-50数据集上进行声学事件检测任务的测试，并和一些其它的CNN模型进行对比，测试结果如表1。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 |

紧接着，我又将其在TUT Urban Acoustic Scenes 2018 dataset上进行声学场景分析任务的测试，并与Baseline和WaveMsNet进行对比，效果优于二者，如表2所述。

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

最后，在说话人识别任务上进行测试，数据集为TIMIT。作为对比，测试了MFCC、FBANK等人工特征和标准CNN结构的错误率，结果如表3。

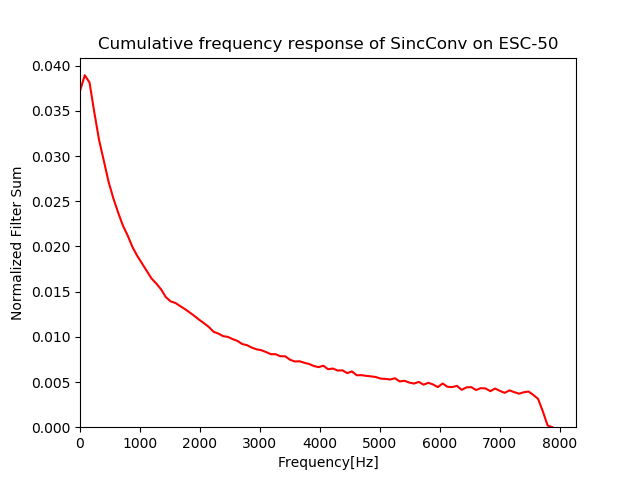
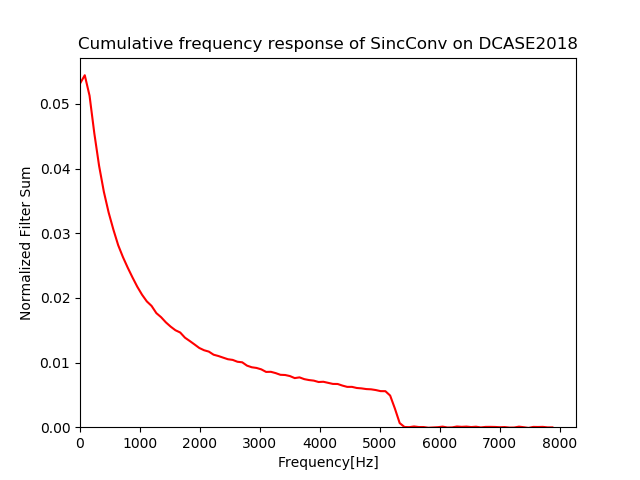
|  |  |
| --- | --- |
| Classification Error Rate of TIMIT Dataset | |
| Model | CRE (%) |
| DNN-MFCC | 0.99 |
| CNN-FBANK | 0.86 |
| CNN-Raw | 1.65 |
| **SCNN (ours)** | **0.87** |

4.3 模型分析

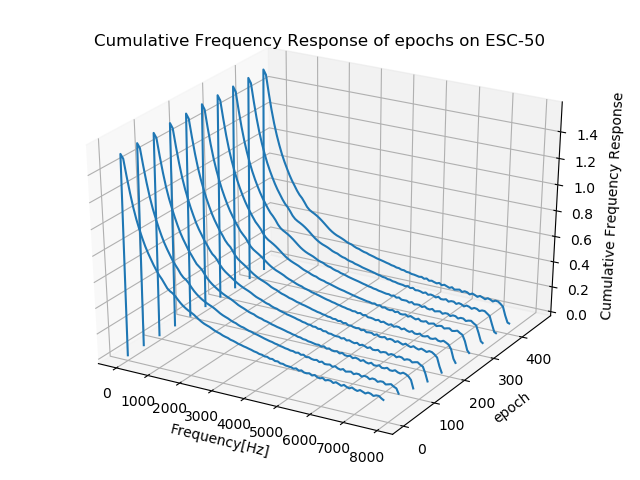
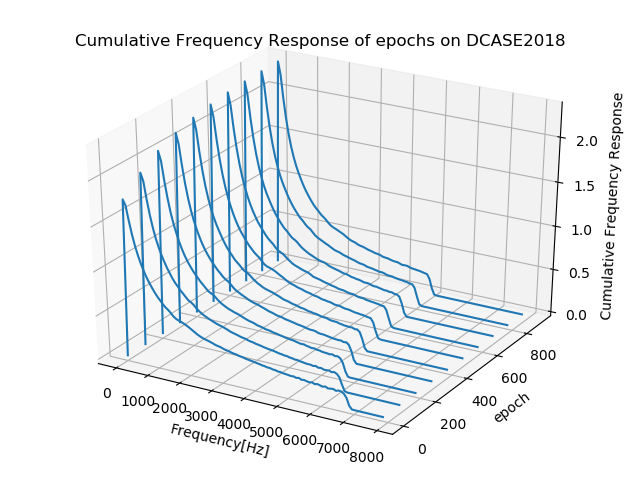
4.3.1 频域响应图

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

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数据集下的累积频域响应图，如图。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。



并且，随着训练的进行，我们可以绘制累计频域响应图的变化过程，进而将滤波器的参数学习过程可视化。



为了与真实数据情况进行验证，随机抽取各个数据集中的几组数据，绘制其原始波形和Log-mel频谱图，进行对比。

4.3.3 基于说话人识别任务的模型分析

4.4 本章总结

第5章 结束语

5.1 本文主要内容梳理

5.2 待解决问题及后续工作

5.3 知识技能学习情况

5.4 职业素养和工程理论的学习和培养

5.5 毕业论文训练收获及体会

参考文献

致谢