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器方法，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)。

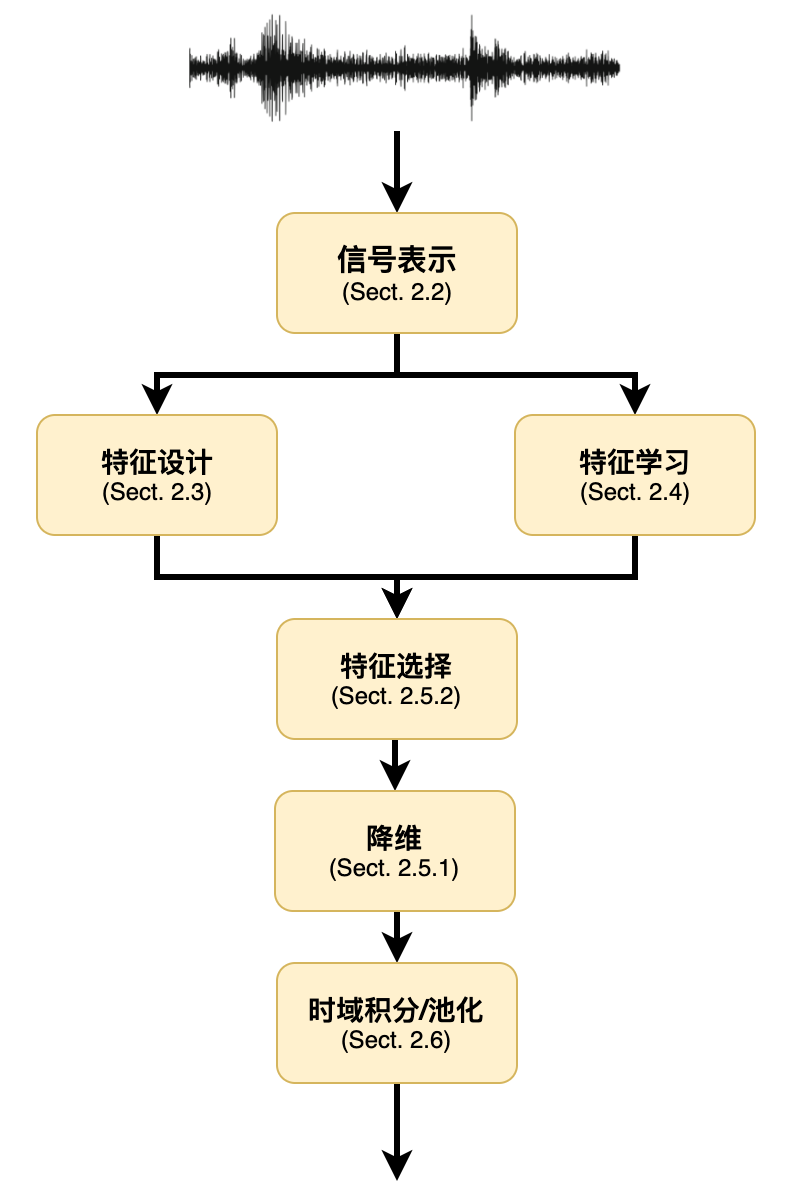


图1

2.2 信号表示

第三章 实验及相关配置

第4章 实验结果分析

第5章 结束语

参考文献

致谢