基于卷积神经网络的音频场景分类方法研究

# 第1章 绪论

## 1.1 课题背景及研究的目的及意义

声音，一种传递信息的媒介。我们能从声音中提取到不少有用的信息。伴随着人工智能技术的发展，借助机器辅助人们从声音中提取出信息成为了人工智能领域一大研究方向。具体来说，音频处理任务中包括自动音乐标记、音乐检索推荐、语音识别、音频指纹、音频场景识别等应用场景。本文主要研究音频场景识别。

音频场景识别的应用场景包括设计上下文感知服务（Adams, Want，1994），智能可穿戴设备（Xu, Li, Lee,2008），机器人导航系统（Chu, Narayanan, Kuo, Matari，2006）和音频归档管理（Landone, Harrop, Reiss，2008）。此外，智能个人助理也是一个受到音频场景识别推动的领域。 智能个人助理是通过分析各种输入数据—包括音频，图像，用户输入或位置，天气和个人时间表等上下文信息，自动进行推荐和执行操作的软件代理。当今智能个人助理服务的代表有Google的Google Now、微软的Cortana、Apple的Siri以及亚马逊的Alexa。这些服务从环境音频中提取上下文信息，可以向用户自动推荐具有价值的信息，是一种极具实用价值的人工智能应用。

音频场景识别属于计算听觉场景分析(CASA)的子领域，其主要目标是通过分析声音使设备能够理解其环境。实现过程为：先对采集到的音频信号进行预处理，再从中提取用于区分环境的有用特征，最后根据这些特征进行分类。

此前音频场景分类基于将通用分类器（高斯混合模型，支持向量机，隐马尔可夫模型）应用于手动提取的特征，例如梅尔频率倒谱系数。近年来，得益于计算机速度的提升与深度学习的快速发展，人们逐渐意识到，可以尝试用深度学习的自动特征提取的特性来代替以往低效的手工提取。正如“深度学习”一词所表明的那样，该方法通过使用非线性模块堆叠多个层来进行低层数据的高级表示。有几种深度学习体系结构的变体，卷积神经网络其中的一种，由于其在学习独特的局部特征方面的优越性能，被广泛用于图像分类、语音识别、自然语言处理。与其他深度学习结构相比，卷积神经网络在图像和[语音识别](https://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3%E8%AF%86%E5%88%AB)方面能够给出更好的结果。相比较其他深度、前馈神经网络，卷积神经网络即使在有限的数据集和简单的数据增强下也可以有效应用于环境声音分类任务。更重要的是，可用数据集规模的显著增加很可能大大提高训练模型的性能。得益于卷积神经网络对数据集的利用程度高及高效的类别学习特性，可以看出卷积神经网络对音频场景分类任务有很高的价值。

## 1.2 国内外研究现状

早在1997年，MIT媒体实验室就已经展开了音频场景识别的工作。在研究起步时期，识别率不甚理想。而随着如今智能设备大量涌现，优秀的计算机性能与深度学习技术的发展共同推动了该领域的研究进程。从过程来看，音频场景识别通常需要进行预处理、特征提取、分类器分类这一系列流程。因此，基于本文的需要，下面分别从音频特征提取、卷积神经网络、音频场景识别三个方面介绍国内外研究现状。

### 1.2.1 音频特征提取研究现状

许多关于音频场景识别的工作都致力于特征提取部分，原因在于良好的特征简化了分类器的设计，而识别能力欠佳的特征很难用任何分类器来补偿。此前受到语音识别领域的推动，有许多特征已经被投入到广泛的研究与应用之中。下面将介绍几种常用特征，并简述其原理：

（1）简单时频特征：(Eronen,2006)及（Malkin,Waibel,2005）等人设计的音频场景分类系统中，采用了这类特征，其特点为可以通过简单时域计算或通过傅里叶变换得到。其中包括过零率：其测量信号内的符号变化的平均速率，并且与单声道声音的主频率相关;光谱质心，测量光谱的质心，它与亮度的感知有关；以及频谱滚降，它识别频率高于设定阈值的频率。

（2）频带能量特征（能量/频率）：(Eronen,2006)等人中使用的这类特征，是通过在指定频带上对幅度谱或功率谱进行积分得到的。得到的系数用来衡量不同子带内存在的能量，并且还可以表示为子带能量与总能量之间的比率，以编码信号中最突出的频率区域。

（3）听觉滤波器组：能量/频率特征的进一步应用在于通过滤波器组分析音频帧，以模仿人类听觉系统的响应。它通过了一组带通滤波器，输出具有一定中心频率的子带信号。其中，（Sawhney和Maes，1997）使用Gammatone过滤器，（Clarkson等，1998）。采用了Mel级滤波器组系数（MFC），而（Patil和Elahili，2002）使用了听觉谱图。

（4）倒谱特征：倒谱特征是为了某些时候便于计算，将原信号的频谱转化为类似分贝的单位，再对其做逆傅里叶变换，将其视为一种新信号处理。MFCC则为音频场景识别中最常用的倒谱特征之一。其得以大量应用的主要原因就在于MFCC的频带划分是在梅尔刻度上等距划分的，它相比于正常对数倒谱中的线性间隔频带更接近人类的听觉系统。在之后的章节中将详细介绍该特征的原理以及提取过程。

（5）发声特征：每当认为信号中包含谐波分量时，可以估计其基频f或一组基频，并且可以定义特征组以测量这些估计的特性。在音频场景识别的情况下，谐波分量可能与音频场景内发生的特定事件相对应，且可以通过其识别确定不同的音频场景。盖格等人。采用与系统中每帧的基频相关的发声功能。 （Krijnders和Holt，2013）提出的方法基于提取音调拟合特征，即从音频信号的感知动机表示导出的一系列发声特征。首先，计算耳蜗图以提供受人耳蜗特性启发的声学场景的时频表示。然后，评估每个时频区域的音调以识别声学场景中的音调事件，从而产生音调拟合特征向量。

（6）线性预测系数（LPC）：线性预测是根据[线性预测](https://zh.wikipedia.org/wiki/%E7%BA%BF%E6%80%A7%E9%A2%84%E6%B5%8B)模型的信息用[压缩](https://zh.wikipedia.org/wiki/%E6%95%B0%E6%8D%AE%E5%8E%8B%E7%BC%A9)形式表示[数字](https://zh.wikipedia.org/wiki/%E6%95%B0%E5%AD%97)[语音](https://zh.wikipedia.org/wiki/%E8%AF%AD%E9%9F%B3)[信号](https://zh.wikipedia.org/wiki/%E4%BF%A1%E5%8F%B7)[谱包络](https://zh.wikipedia.org/w/index.php?title=%E8%B0%B1%E5%8C%85%E7%BB%9C&action=edit&redlink=1)。它是最有效的音频分析技术之一，也是低位速下编码方法高质量语音最有用的方法之一，它能够提供非常精确的语音参数预测。这类特征已被用于分析被模拟为自回归过程的语音信号。在自回归模型中，给定时刻t的信号s的样本表示为L个先前时刻的样本的线性组合：

(1)

其中组合系数确定了模型参数和是一个剩余项。 LPCS的值与建模信号的频谱包络之间存在映射，因此编码关于声音的一般频谱特性的信息。 Eronen等人在他们提出的方法中采用了LPC特征。

### 1.2.2 卷积神经网络研究现状

卷积神经网络（Convolutional Neural Network, CNN）深度学习中一种重要的网络架构。它是一种[前馈神经网络](https://zh.wikipedia.org/wiki/%E5%89%8D%E9%A6%88%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C" \o "前馈神经网络)，它的人工神经元可以响应一部分覆盖范围内的周围单元。卷积神经网络由一个或多个卷积层和顶端的全连通层（对应经典的神经网络）组成，同时也包括关联权重和[池化](https://zh.wikipedia.org/w/index.php?title=%E6%B1%A0%E5%8C%96&action=edit&redlink=1" \o "池化（页面不存在）)层（pooling layer）。这一结构使得卷积神经网络能够利用输入数据的二维结构。

卷积神经网络结构的确立源自于1997年（Le Cun等，1997）人的论文，他们设计了一种名为LeNet-5的多层人工神经网络，可以对手写识别数字做分类。卷积神经网络也像其他神经网络一样可以使用反向传播算法进行训练。卷积神经网络的热潮的掀起则是由于2012年开始举办的ImageNet图像分类比赛。赛中（Krizhevsky等，2012）提出了一个经典的卷积神经网络结构，并在图像识别任务上取得了重大突破。其方法的整体框架叫做 AlexNet，与 LeNet-5类似，但层次结构上要更加深一些。同时使用了非线性激活函数ReLu[5]与Dropout[6]方法，取得了卓越的效果。

### 1.2.3 音频场景识别研究现状

Sawhney和Maes在1997年MIT媒体实验室的技术报告中提出一种专门解决ASC问题的方法。作者记录了一组包括“人”，“声音”，“地铁”，“交通”和“其他”的一组数据集。他们利用语音分析和听觉研究借鉴的工具从音频数据中提取了几个特征，采用递归神经网络和k最近邻标准对特征和类别之间的映射进行建模，并获得68％的整体分类准确率。一年后，来自同一机构的研究人员Clarkson等通过戴着麦克风录制连续的音频流，同时进行一些超市自行车旅行，然后自动将音频分割成不同的场景（如“家”，“街道”和“超市”）。他们将从音频流中提取的特征的经验分布拟合成隐马尔可夫模型（HMM）。

与此同时，实验心理学的研究则着重于理解驱动人类对声音和场景进行分类和识别的能力的感知过程。 Ballas发现识别声音事件的速度和准确性与刺激的声学性质、它们发生的频率及是否它们可以与物理原因或声音刻板印象相关联有关（Ballas，1993）。佩尔顿等人（Peltonen等，2001）观察到人类对音频场景的认识是通过识别典型声音事件（如人声或汽车发动机噪声）来实现的，并且确定了人类识别25个声场中的能力的整体准确率为70％。 Dubois等人（Dubois等，2006）研究了在不是实验者先验的情况下，个体如何定义他们自己的语义类别分类。最后，Tardieu等人（Tardieu等，2008）测试了语义类的出现以及在火车站范围内对声场的识别。他们在报告中说，声源、人类活动以及房间效应（如混响）是促成音频场景形成的因素，也是类别为固定先验情况下的识别线索。

受心理声学/心理学文献的影响，这些文献强调音频场景识别的局部特征和全局特征，一些麻省理工学院研究人员则侧重于音频的时域特征。 Eronen等人（Eronen等，2003）采用Mel频率倒谱系数（MFCCs）来描述音频信号的局部频谱包络，用高斯混合模型（GMMs）来描述其统计分布。然后，他们通过利用训练信号种类的知识的判别式算法来训练HMM，以解释GMM的时域演变。 Eronen及其合作者通过考虑更多的特征，和在分类算法中增加一个特征变换步骤，进一步推进了这项工作，在18种不同的声场中获得了总体58％的准确性。

尽管关于ASC系统的文献丰富，但研究界缺乏协调一致的标准来评估和测试解决这个问题的算法。2013年，IEEE音频和声学信号处理（AASP）技术委员会首次组织了DCASE (音频场景和事件检测和分类)挑战赛，以测试和比较ASC和事件检测与分类算法。这一举措符合信号处理领域旨在促进可再生研究的目标。过去几年来，本挑战赛中已经提出了许多音频处理技术，对整个ASC系统的发展做出了极大的贡献。

Santoso等人（Santoso等，2016）使用卷积神经网络网络来构建分类器。该系统是基于计算机视觉领域工作中采用的体系结构而设计的，具体来说，用来构建本系统分类器的结构为NIN（网络中的网络）（Lin等，2013）。NIN体系结构被提出来改善局部模型在CNN卷积层的抽象能力。 NIN用一个更有效的非线性逼近器替代了对CNN中的数据补丁进行抽象的模型。 在NIN架构中，抽象模型被MLP网络取代。此外，NIN架构取代了传统CNN的分类方法。 在CNN架构中，特征映射被连接到作为分类器的传统MLP网络。 NIN架构使用全球平均汇集来取代这种分类方法。 NIN架构直接使用最后一个卷积层中的特征映射来构建分类器。 该体系结构取得特征映射的平均值，并将生成的向量直接输入到softmax层。在本系统的特征提取部分，使用梅尔频谱系数（MFCC）作为分类器的输入向量。分类器使用来自MFCC特征集的每个帧进行训练，然后对每个帧的结果进行阈值化并投票选择音频数据的最终场景标签。其系统准确度胜过DCASE挑战的基准系统。系统平均准确率为78.83％，基准系统平均准确率为72.57％。

辛德勒等人（Schindler，2016）通过使用CQT（常数Q变换）特征作为CNN的输入来增强结果。CQT是一个时频表示，其所有频段的Q因数,即中心频率与带宽的比值相等。 CQT本质上是一种小波变换，这意味着对于低频率，频率分辨率更好，时间分辨率对于高频率更好。使用CQT的动机来自于音乐感知领域的观点：人类听觉系统在大部分可听频率范围内近似为“常量Q”。该系统的关键是利用CQT以足够的分辨率捕获来自低频和高频的基本音频信息，并创建一个并行CNN架构，该架构能够及时捕获这两种频率。所呈现的深度神经网络架构已经比DCASE 2016声场景分类任务组织提供的基线系统超出10.7％的相对改进，在开发集合上达到80.25％。此外，它在评估集中达到了83.3％，在DCASE 2016挑战任务中排名第35。

由于此前在音频场景分类领域缺乏大型标记的声音数据集，获得这些数据集通常既昂贵又不明确。来自MIT的Aytar等人寄希望于通过利用视觉和声音之间的自然同步来学习来自未标记视频的音频特征来扩大规模，因此他们利用超过一年的野外采集的声音来学习语义丰富的音频特征（Aytar等，2016）。未标记的视频可以大规模、低成本的获得，且具有音频信号。计算机视觉方面的最新进展使机器能够高精度地识别图像和视频中的场景和对象。而如何将视频中的知识转化为标记音频的标签成为了研究的关键。在实验中，他们使用了可以直接在原始音频波形上学习的卷积神经网络，通过将知识从视觉传输到声音进行训练。尽管网络是通过视觉监督进行训练的，但网络在推理过程中不依赖视觉。结果表明，与简单的全连接的网络或较早的图像分类体系结构相比，最先进的图像网络在音频分类方面具有出色的结果。其对较大的标签集词汇进行训练可以提高性能，尽管在对较小的标签集进行评估时性能稍有提高。

## 1.3 研究内容

通过之前的总结我们可以发现，此前应用的一般音频场景识别方法如GMM，HMM等存在着对音频集利用不充分的问题。根据之前的经验，本文采用卷积神经网络作为分类器，以提高音频集的利用效率。

具体研究内容章节如下：

# 第2章 音频场景分类基线系统

## 2.1系统结构简介

本章主要内容为介绍基于MFCC特征及GMM分类器的音频场景分类基线系统。MFCC作为语音识别领域的重要特征，有着良好的区分性能。而且研究也表明（Reynolds等，1995），在GMM作为分类器与MFCC等性能良好的特征配合时，有着出色的分类性能。因此选取以上经典方法作为音频场景分类的基线系统，以判断之后提出的基于卷积神经网络的分类器的优势。

本文中的基线系统主要包括五个部分，分别为音频数据集、特征提取与处理、系统训练、系统测试、系统评估

## 2.2特征处理

### 2.2.1 MFCC特征

提取音频信号的最佳参数表示是产生更好识别性能的重要任务。这一阶段的效率对下一阶段很重要，因为它会影响之后的分类效率。

人类之所以能在复杂的声音环境中判断出不同的环境，主要在于耳蜗的功劳。耳蜗可以看作为一个滤波器组，帮助人们过滤20-20KHz的音频。问题在于耳蜗对于听觉范围内频率的灵敏度并不是线性的，存在一种映射关系。因此为了模拟出人耳的频率响应，（Davis等，1980）提出了MFCC特征。

MFCC的整个过程如图2所示：

****

如图2所示，MFCC由七个步骤组成。下面将具体介绍这些步骤：

（1）预加重

许多现实中的音频信号，它们的功率谱随频率的增加而减小，其大部分能量集中在低频范围内。这就造成了音频信号中高频的信噪比可能降到不能容许的程度。预加重的目的是更加凸显音频信号的高频特性。通常借助高通数字滤波器来实现。

设第n时刻的语音采样值为x(n)，经过预加重处理后的结果是:  
其中a为预加重系数，通常取0.9~1.0之间。

（2）分帧

分帧将从模数转换（ADC）中得到的音频样本分割为长度在20至40毫秒范围内的小帧。音频信号被分N个样本的帧。相邻帧由M（M<N）分隔。通常取M=100，N=256。

（3）加窗（海明窗）

在预加重与分帧完成后，需要为每一帧加上海明窗（Hamming window）。加窗是为了控制数据处理量，每次仅处理窗中数据。由于之后FFT中的处理对象是有限长信号，对无限长信号强行进行有限个点的FFT会丢失频率信息导致频谱泄露。海明窗的公式可以总结如下：

假设窗函数为：。且为每帧中的样本数，为输出信号，为输入信号，则加窗结果如下：

其中，

（4）快速傅立叶变换（FFT）

将N个样本的每个帧从时域转换到频域。具体过程如下：

（5）Mel滤波

FFT频谱中的频率范围非常宽，语音信号不遵循线性标度。因此通过如图3所示的Mel标度滤波器组。



该图表示了一组三角形滤波器，用于计算滤波器频谱分量的加权和，使得处理后的输出近似于Mel标度。每个滤波器的幅度频率响应是三角形的，在中心频率处等于1，在两个相邻滤波器的中心频率处线性减小到零，且每个滤波器输出是其滤波后的频谱分量的总和。最后，使用以下等式计算给定频率f的Mel：

（6）离散余弦变换

该过程为使用离散余弦变换（DCT）将log-Mel频谱转换为时域的过程。转换的结果称为Mel频率倒谱系数。系数集称为声矢量。因此，每个输入都被转换为声学矢量序列。

（7）差分计算

以上取得的倒谱参数只能反应音频信号的静态特性。为了提高信号的识别性能，应采用音频信号静态特性的差分谱来描述音频信号的动态特性。因此引入了13个一阶差分特征（12个倒谱特征加能量），以及39个二阶差分特征。从时间t1到t2的窗口中信号x的帧能量如下式所述：

13个一阶差分特征中的每一个特征表示MFCC特征中对应的倒谱或能量特征的帧之间的变化，而39个二阶差分特征中的每一个表示对应的一阶差分特征中的帧之间的变化。其中一阶差分的计算如下式所述：

其中，表示时刻的倒谱系数。

## 2.3 GMM模型原理

# 第3章 基于MFCC 和CNN的音频场景识别系统

# 第4章 基于MFCC 和CNN的音频场景识别系统改进

# 结 论

# 参考文献

Eronen A J, Peltonen V T, Tuomi J T, et al. Audio-based context recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2006, 14(1):321-329.

Malkin R G, Waibel A. Classifying user environment for mobile applications using linear autoencoding of ambient audio[C]// ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing. 2005.

Krijnders J, t Holt G. Tone-fit and MFCC scene classification compared to human recognition[J]. Energy [dB], 2013, 400(450): 500.

Sawhney N, Maes P. Situational awareness from environmental sounds[J]. Tech-nical Report, Massachusetts Institute of Technology, 1997.

Clarkson B, Sawhney N, Pentland A. Auditory context awareness via wearable computing[J]. Energy, 1998, 400(600): 20.

Patil K, Elhilali M. Multiresolution auditory representations for scene classification[J]. cortex, 2002, 87(1): 516-527.

Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

Ballas J A. 1993. Common factors in the identification of an assortment of brief everyday sounds[J]. Journal of experimental psychology: human perception and performance, 19(2): 250.

Peltonen V T K, Eronen A J, Parviainen M P, et al. 2001. Recognition of everyday auditory scenes: potentials, latencies and cues[J]. PREPRINTS-AUDIO ENGINEERING SOCIETY.

Dubois D, Guastavino C, Raimbault M. 2006. A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories[J]. Acta acustica united with acustica, 92(6): 865-874.

Tardieu J, Susini P, Poisson F, et al. 2008. Perceptual study of soundscapes in train stations[J]. Applied Acoustics, 69(12): 1224-1239.

Eronen A, Tuomi J, Klapuri A, et al. 2003. Audio-based context awareness-acoustic modeling and perceptual evaluation[C]//Acoustics, Speech, and Signal Processing, 2003.

Lin M, Chen Q, Yan S. 2013. Network in network[J]. arXiv preprint arXiv:1312.4400.

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.

Lidy T, Schindler A. 2016. CQT-based convolutional neural networks for audio scene classification[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016). DCASE2016 Challenge, 90: 1032-1048.

Aytar Y, Vondrick C, Torralba A. 2016. Soundnet: Learning sound representations from unlabeled video[C]//Advances in Neural Information Processing Systems. 892-900.

Santoso A, Wang C Y, Wang J C. Acoustic scene classification using network-in-network based convolutional neural network[R]. DCASE2016 Challenge, Tech. Rep, 2016.

Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models[J]. IEEE transactions on speech and audio processing, 1995, 3(1): 72-83.

# 致 谢