

**文献综述报告**

**（ 2021 届本科 ）**

**学 院**：  **信息学院**

**专 业： 软件工程**

**班 级： 17软工2班**

**姓 名：**  **骆春江**

**学 号：**  **1759231**

**指导教师： 王甘霖**

**基于深度学习的音频分析处理研究**

**摘要：**

在声音信号检索（Sound Information Retrieval, SIR）领域，音频场景分类（Environmental Sound Classiﬁcation, ESC）作为该领域的热点问题，致力于通过分析从各种音频信号中提取的复杂特性，识别其对应的特定场景所包含的语义标签，从而对其周围环境进行感知和理解，最终实现特定音频场景的分类。

常用音频信号特征提取方法为梅尔频率倒谱系数（MEL Frequency CepstrumCoeﬃcient, MFCC)。这种方法虽然抗干扰能力强，能够抓取音频数据中最有辨识度的部分，但却只能分析信号的短时特征，往往不足以完整刻画整个音频数据的结构特点。近年来深度学习技术日益成熟并作为最有效的特征提取方法之一，已在机器学习、图像识别、自然语言处理等诸多领域取得突破性进展[1]。卷积神经网络（Convolutional Neural Networks, CNN）作为典型的深度学习网络框架，具有权值共享和局部连接等特性，特别是带有池化层的卷积神经网络在对城市声音声源分类方面卓有成效。然而，池化操作往往会导致信息的大量丢失，从而影响分类结果准确性。

本文在 CNN 具有良好结构特征分析能力的基础上，将传统音频信号特征提取方法 MFCC 进行了进一步结构特征分析，探索更好的深度学习方法以解决传统的音频场景分类问题。首先通过深度学习的经典模型 CNN 进行实验，并采用了结构特殊的膨胀卷积方法对比，发现膨胀卷积由于其“网格型”结构可以在原始参数不增加的情况下，扩大感受野范围，覆盖更多帧，从而很好代替传统带有池化操作的卷积运算。同时，通过对膨胀卷积结构特点的进一步研究，发现膨胀系数变大或扩大膨胀卷积层数将使实验精度降低。认为在膨胀卷积模型中存在固有“网格”缺陷，因此会丢失很大信息，过度放大的感受野使得框架太大而不能获得声音信号随时间变化的特性。可以预见，基于深度学习的音频场景分类问题在今后的工作中还存在很多值得深入探究的内容。

本文主要研究内容及取得成果如下：

（1）整理总结国内外音频信号处理问题及深度学习技术的发展现状，发现传统音频信号处理只能分析信号的短时特征，后续识别问题步骤繁多复杂，且主要基于一般浅层分类器的应用。通过研究典型深度学习方法，寻找适合的结构化模型在语音识别分类上进行实践应用。

（2）深度学习的方法多种多样，不同结构对不同场景特征敏感度不同，识别性能存在差距。本文研究了带有池化操作的传统卷积神经网络的音频场景特征提取分类方法。并在模型设计中引入了膨胀卷积思想，通过这种特殊结构的卷积操作形式，在城市声源数据集中取得了比传统卷积神经网络更好的结果。

（3）深入研究膨胀卷积的结构对实验结果的影响，发现扩大膨胀率或膨胀卷积层数将使实验分类精度降低。这可能是音频信号具有短时稳定性，膨胀模型有“网格”连接缺陷，经过 MFCC 处理的特征再通过这种网格结构，覆盖帧的范围严重改变，最终影响整体音频信号的特征提取。

**关键词：**深度学习；音频场景分类；卷积神经网络；膨胀卷积

**研究背景和意义**

**（1）音频信号检索实用性强、涉及面广、应用领域广泛**

声音信号是多媒体信息中重要的组成部分，包含语音、语言、音乐等多种类别，往往承载大量有价值的信息和线索，是人类感知事物的重要形态。通过音频信号传递信息是生物体最有效、最直接、最重要、最方便也是最常用的信息交流方法。近年来，随着多媒体技术的快速发展和普及，以及声音采集技术和互联网技术的快速发展，每天都会有大量的音频数据产生，且数据信息的保存形式多种多样。由于音频信息直接与社会生活、工业生产部门等行业的语音咨询与管理、声音的控制等各种实际应用领域相接轨。例如行政办公、交通网络、金融管理、社会公安、商业、旅游、医疗服务等，对音频信号的研究与探讨一直以来得到各行各业的广泛关注。随着信息化时代的到来，计算机方向发展走向便携化，计算环境也日趋复杂，人们希望通过对现代化的技术和高质量的学习方法研究，更有效地对音频信号进行处理，并广泛应用到机器学习任务中。

音频信号检索（SIR）作为一个热点问题在众多应用领域被人们广泛讨论[2]，除了其实用性，还有一个重要的原因就是这门学科始终与当前信息科学领域最重要的前沿课题保持密切的联系。而且，这门学科与心理学、生理学、计算机技术、人机交互、模式识别、人工智能等都有着密切的联系，是一门涉及面广泛、实用性强的综合性交叉学科。在音频检索领域[3]，研究者一直在尝试采用更高效更适合的建模方法对一系列语音语言信号进行建模，并希望能够通过对这一特殊信号的研究，使其广泛应用到社会生活的各个方面。例如自然语言处理[4]，不同风格或者不同乐器的音乐生成，音乐风格或者作曲家分类等[5–7]。通过对现代化的技术和高质量的学习方法研究，使人们能够更有效地对音频信号进行处理，并广泛应用到机器学习任务中，对促进社会发展，提高人们对声音更深层次的理解和应用具有十分重要的意义。

**（2）音频场景分类问题应用前景丰富，有很高研究意义与价值**

在这些海量的音频数据当中[8]，除了一些含有具体实际意义的语言语音类音频信息外，还有一些场景类音频信号也是其中重要的组成部分。音频场景分类（ESC）问题就是对这种非语音的环境场景声音，通过有效的分析与感知，进而进行高层次语义分析的研究方向。这种技术指的是将语义标签与音频流关联起来，从而识别出它所生成的环境的一种任务。首先从各种音频信号中提取复杂特性，通过拆解、分析、归类等方法，识别其对应的特定场景所包含的语义标签，从而对其周围环境进行感知和理解，最终实现特定音频场景的分类。音频场景的识别分类有很多潜在应用前景，比如盲人可以通过这种技术准确了解自己所处的环境，辨别自己所在的具体位置；在军事领域、刑侦领域，这种技术可以与视频监控相辅相成，比如在视频质量不佳情况下，通过该技术的运用，可以帮助刑侦人员更好地监控与定位，且相比于视频监控，这种方法需耗费的资源相对较少，所需要的设备也不用特别大，可移动性强；在智能机器人方面，音频场景分类技术对人机交互技术[9]也有很大推进作用，它可以使我们的机器人更加的“聪明”，“耳朵”更加“灵敏”，辅助机器人及时了解周围复杂多变的场景，判断所处环境，明确自己所处的位置，从而实现下一步的有效应对。

鉴于音频场景分类问题广泛的应用前景，国内外学术界、工业界及政府部门、军事部门对该问题给予了相当大的关注，一些专门针对音频信号处理的会议及测评相继展开。例如欧盟 CHIL 人机交互项目和美国国家标准技术研究所举办的CLEAR 评测，主要针对各种音频类型及类型之间的关系进行研究分析。从 2005 年开始，由国际音乐信息检索系统评估实验室主办的年会 ISMIR，就音乐信号的研究每年召开一次会议。这些国际测评工作及会议组织强有力的推动了音频信号处理的研究发展。在国内，针对数字信号处理、语音信号分析领域的音频特征提取工作也得到了各类国家重点基金和企业等的大力支持。由此可见，音频场景分类问题在理论和实际应用中都有着很高的研究意义和价值。

**（3）深度学习方法有望解决传统音频分类方法及浅层学习存在的挑战**

传统的音频分类方法往往将原始数据经过短时傅里叶变换形式或者采用基于倒谱域的梅尔频率倒谱系数（MEL Frequency Cepstrum Coeﬃcient, MFCC）[10]提取特征序列，以帧（frame）为单位，利用混合高斯模型（GMM）[11]、隐马尔科夫模型（HMM）[12]、或者 GMM-HMM 模型[13]进行进一步的训练。但是这些方法不仅受限于高度预处理的声学特征分析，还由于这些特征往是短时特征或者仅仅是长时特征的一个统计值，要么很难对音频场景进行完整描述，要么长时间的统计特性会导致信号的局部结构性特性信息丢失，最终很难有效解决复杂的场景音频流分类问题。

以前，绝大多数的机器学习和信号处理方法都采用浅层学习的方法（例如HMM 模型），这种情况局限在样本和计算单元的个数，不利于复杂函数的表现[14]，面对复杂的分类问题或复杂的信号（比如自然图像、人类或场景的语言、声音等）其建模能力、泛化能力就会受到很大的制约。2006 年，机器学习领军人物、多伦多著名教授 Geoffrey Hinton 在《Science》上发表文章[15]，由此在学术界和工业界深度学习引起了广泛的关注[14]。作为机器学习的一个新的研究领域，深度学习通常也被叫做深层结构学习或分层学习，目的在于模拟人脑的机制，建立类似人脑分析学习的神经网络来对数据进行分析解释。深度学习采用一种基于无监督特征学习和特征层次结构的学习方法，通过学习一种深层的非线性网络结构，实现复杂函数逼近，表征输入数据分布式表示，展现了从少数样本中集中学习数据以及其本质特征的强大能力[14,16]。由深度学习发展而来的技术对人工智能、信息信号处理等各个方面产生了深远影响，在图像处理[17–19]、音频信息检索[20,21]等领域也有极大突破[8]。卷积神经网络是深度学习方法中较为独特的一个，类似生物学感受野（Receptive Field）机制，利用局部感受野对整个研究范围进行逐层覆盖；通过权值共享结构，降低网络模型复杂度，减少整体权值数量；采用时间或者空间的子采样结构，更好的吻合了输入信息和网络拓扑结构，并且可以获得某种程度的位移、尺度、形变鲁棒性，在图像和语音识别[22,23]问题方面具有独特优势。

综上，基于深度学习的方法在音频场景分类问题上研究十分有必要。通过探寻不同的网络结构与模型，有望实现该领域更大的突破。本文在基于卷积的深度学习基础上，对音频分类问题展开进一步的研究，引入了一种特殊的卷积结构（膨胀卷积）对城市场景音频进行分类实验。通过实验，分析了模型的结构特点以及针对音频信号该模型主要影响实验效果的因素，对未来继续深入挖掘音频数据的特点以及将深度学习方法与音频数据处理的方法有效结合有一定借鉴意义。

**国内外研究现状**

**音频场景分析研究现状**

自然界的生物体往往从听觉、视觉、触觉等多个角度对事物进行感知分析。声音是人类常用的感知工具，承载着大量有价值的信息和线索，通过音频传递信息是人类最有效、最直接、最重要也是最常用的信息交流方式。随着音频采集技术及互联网技术的高速发展和普及，作为多媒体应用中重要的组成部分，每天都会有大量的音频数据产生。机器听力系统执行类似人类听觉系统的任务，这是机器学习、机器人和人工智能等领域广泛研究课题的一部分。日常音频事件通常提供特征信息的是环境背景音，不仅仅包括音乐或语音数据，而且它们的结构往往更加多样化和令人困惑。

声音信号检索（SIR）作为一个热点问题在很多应用领域被人们广泛讨论。例如在海洋哺乳动物声音分类问题中[24]，一个新的海洋哺乳动物分类计算模型被提出。通过该模型对在线海洋动物声音数据库的数据进行特征提取及分类，科学家们可以更准确探测、识别和定位不同的濒危物种以及可能对海洋生态系统造成伤害的高强度人为来源。在飞机分类领域，飞机起飞的声学影响是监测和研究的重要课题。分析基于起飞特性产生高水平噪声的飞机的类型或类别，提出一些关于机场附近区域的飞机分类和声学影响的估计方法，可以更好的识别具有性能的飞机类别，从而更深入地分析飞机噪声。例如采用基于 MFCC 和倍频程分析进行特征提取，将神经网络分类模型用在真实环境中记录飞机起飞信号[25]，根据安装的发动机类型，就可以将飞机分为几类。此外基于起飞噪声信号分割的飞机类识别多层神经模糊模型[26]，可以通过机场噪声监测，实时对环境噪声测量，对机场起飞噪声进行更好评估以及智能数字处理，最终还通过神经模糊模型可以更准确识别飞机类别。

在嘈杂城市环境中，车辆声、人群声、汽车喇叭声等各种噪声混合在音景中，而且结构复杂多样，导致城市音频分析也变得特别困难。城市环境中声音事件的分类有很多应用，包括上下文感知计算[27]、监视[28]，或者基于内容的多媒体检索技术，比如城市应用中的高亮提取[29]和视频摘要[30]，以此来确定全市的重要事件等。更重要的是，通过对城市声音和噪声模式的理解，也有可能提高城市居民生活质量，这在一定程度上通过配备多媒体传感器网络[31]的智能城市发展实现。例如一种基于 MATLAB 和 BSS 的环境声学分析改进方法[32]实现了有效从记录的混合声源中分离噪声源，并为改善城市地区常见噪声源监测和分析结果提供了依据。还有一种基于源特定字典的混合声音主要环境噪声分类方法，对主要声源进行分类[33]，成功识别了信号中同时存在的两个，三个和四个源类间混合音频中的主要来源。有利于对产生最高噪声水平的源进行分类从而减少城市区域的噪声污染。此外，它还确定了可以控制或隔离的来源，以降低影响人口健康和舒适的噪音水平。

在声音信号检索领域（SIR），音频场景分类（ESC）是指根据音频流的声学特征，将特定的语义标签与音频流关联起来，以识别音频流产生的特定场景环境。使设备能够通过分析声音来了解其所处的环境是机器听力研究的主要目标，这是一个广泛的研究领域，最早针对音频场景识别问题提出的学术性文章可以追溯到 1997年 MTT 实验室[34]。区别于传统心理学研究方法，这种方法不完全致力于理解人类感知的音频场景识别机制，而是更多依赖信号处理技术以及一些人工智能和机器学习的方法手段，实现音频场景的自动识别和分类。然而，这些工作仍然按照传统的模式识别框架进行，即音频场景数据特征分析、提取（比如频谱特征），最后采用适合的分类模型进行分类，例如支持向量机（SVM）、隐马尔科夫模型（HMM）等一般分类器的应用。近年来，在这种背景下，深度学习技术缓慢引入[35,36]，详细分析了最常用的方法，但仍然主要局限于分析高度预处理的声学特征[37,38]。

**深度学习研究现状**

传统的模式识别方法往往通过传感器获取数据，首先根据需求将数据进行适当的预处理，然后提取其特征，进行相应的特征选择，最后再进行具体推理、预测和识别几个过程。良好的特征表达是最终提升算法准确性的关键，识别系统的主要计算和测试工作耗时主要集中在特征提取部分。特征越多，给出的信息就越多，识别的准确性就会提高很多。但是特征的样式一般采用人工提取的方式，这种方法费时费力，且很大程度靠以往的经验和运气。而且随着特征增多，计算的复杂度也会增加，探索的空间会变大，在每个特征上可以用来训练的数据就会稀疏。所以，需要有多少特征需要由学习决定。

机器学习的第一次浪潮称为浅层学习[39]。传统机器学习和信号处理技术探索仅限单层非线性变换，且只含单个将原始输入信号转换到特定问题空间的特征结构。这种结构在样本和计算单元有限的情况下及其受限，对复杂的函数表现能力也受到很大的约束，在解决复杂分类问题时泛化能力明显不足。典型的浅层学习结构包括 SVM、高斯混合模型（GMM）、隐马尔科夫模型（HMM）、条件随机场及仅含单隐层的多层感知机、最大熵方法等。

深度学习是机器学习的第二次浪潮[15]。2006 年，Geoffrey Hinton 将深度学习作为机器学习研究的一个新领域，提出深度置信网络（Deep Belief Network, DBN）[15]。由此，在过去的几年里，不仅在传统领域，一些由深度学习引发的科学技术对人工智能、信息处理、信号处理等各个方面都产生了深远影响。深度学习的主要思想是模拟人脑的机制，建立类似人脑分析学习的神经网络来解释例如图像，声音和文本数据等[40]。通过将低层特征组合形成属性类别或更加抽象的高层表示，从而发现数据的分布式特征表示[14]。例如含有多隐层的多层感知机（也称为深度神经网络 DNN）或前馈神经网络。Hinton 提出的深度置信网络就是由一组限制玻尔兹曼机（HBMs）堆叠而成，属于深度生成式模型，其时间复杂度与网络大小和深度呈线性关系，核心部分采用贪婪、组成学习算法，可以最优化网络的权重。还有一种使用 DBN 去初始化 DNN 的操作，被专家称为 DBN-DNN[41]。

深度学习采用分层的多层非线性信息处理方法，根据其研究应用领域的不同可以分为三大类：无监督（生成式）学习的深度网络、有监督（判别式）学习的深度网络以及混合（同时包含生成式模型和判别式模型）的深度网络。无监督学习不采用如目标类别标签的监督信息进行学习，大多通过从网络中采样并有效生成样本，是一种生成式模型。典型的生成式模型有受限玻尔兹曼机（RBM）、深度置信网络（DBN）、生成对抗式神经网络（GAN）和广义除燥自编码器[42,43]。但是也有些例如稀疏编码网络及原始的深度自编码器这些在采样时并不容易，因而本质不能说成生成式的，而且这些模型不一定是概率模型。

在针对语音建模上，递归神经网络（Recurrent Neural Network, RNN）属于有监督（和无监督）学习的深度网络且结果不错。在无监督学习模式下，学习过程中并不涉及类别信息，而是根据先前的数据样本预测未来数据序列，非常适用于例如语音、文本等的建模，但其很难捕捉长时相关性。以 Hessian-free 优化方法训练的 RNN[44]在一定程度上解决了这个问题。在语音识别领域，科学家们开始逐渐探索利用人类发声机制构建动态和深度结构的概率生成模型，将浅层 HMM 模型和表示不同层级的自然语言高级结构结合[45]，并采用用于大规模语音识别和理解的标准统计方法，这种结合的层次模型可以被视为一种深度生成结构。当然，RNN 也可以当成判别模型，此时其输出变成了一组和输入数据序列相关联的标签序列。但是这种方法在解决语音问题上效果有限。之后研究者开始设想把长短时记忆（Long-Short-Term Memory, LSTM）方法嵌入到 RNN 中，让 RNN 无需对训练数据预分段和对输出数据后处理，本身就可以用于序列分类。这种方法在小型语音处理任务中得到了验证[46]。

20 世纪 80 年代，卷积神经网络（CNN）出现[47–49]，成为各种对象分类任务或模式识别的一种有效方法。Krizhevsky 等人的工作[50]标志着 2012 年大规模视觉识别[51]的转折点。从那时起，CNN 取代了依赖人工提取处理特征的技术，在许多模式识别任务上取得了显著的进展，包括交通标志分类、手写数字、行人检测[52]和电子显微镜图像处理[53]。卷积架构虽然主要用于视觉识别上下文，但也成功地应用于语音[54–56]和音乐分析[57,58]。这些工作表明，使用数据局部性特征可以为遇到的问题提供可行的解决方案。那么是否可以将这种深度学习架构用于解决音频场景分类问题呢？2015 年，Karol J. Piczak 等评估了卷积神经网络在解决短小的音频片段在场景分类问题上的潜力[59]，通过使用由两个包含最大池化层的卷积层组成的深度网络以及两个全连接层，在具有增量的音频数据的低级表示上进行训练。与其他特征学习方法相同，即使数据集有限，CNN 模型也能够在音频场景分类问题中显示出功能。

参考文献：

[1] 李祖贺, 樊养余. 基于视觉的情感分析研究综述 [J]. 计算机应用研究. 2015,32(12):3521–3526.

[2] 胡振, 傅昆, 张长水. 基于深度学习的作曲家分类问题 [J]. 计算机研究与发展.2014, 51(9):1945–1954.

[3] 李晨, 周明全. 音频检索技术研究 [D]. Ph.D. thesis, 2008.

[4] Ellen M. Voorhees. Using WordNet to Disambiguate Word Senses for TextRetrieval [C]. Proceedings of the 16th Annual International ACM-SIGIR Con-ference on Research and Development in Information Retrieval. Pittsburgh,PA, USA, June 27 - July 1, 1993. 1993, 171–180.

[5] Tao Li, George Tzanetakis. Factors in automatic musical genre classiﬁcation ofaudio signals [C]. 2003 IEEE Workshop on Applications of Signal Processingto Audio and Acoustics (IEEE Cat. No. 03TH8684). IEEE, 2003, 143–146.

[6] J Stephen Downie. The music information retrieval evaluation exchange(2005–2007): A window into music information retrieval research [J]. Acousti-cal Science and Technology. 2008, 29(4):247–255.

[7] Maximos A. Kaliakatsos-Papakostas, Michael G. Epitropakis, Michael N. Vra-hatis. Musical Composer Identiﬁcation through Probabilistic and FeedforwardNeural Networks [C]. Applications of Evolutionary Computation, EvoAppli-cations 2010: EvoCOMNET, EvoENVIRONMENT, EvoFIN, EvoMUSART,and EvoTRANSLOG, Istanbul, Turkey, April 7-9, 2010, Proceedings, Part II.2010, 411–420.

[8] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, Paul Lamere. Themillion song dataset [J]. 2011.

[9] Jianzhi Jin, Kaihua Xu, Naixue Xiong, Yuhua Liu, Guoqiang Li. Multi-indexevaluation algorithm based on principal component analysis for node impor-tance in complex networks [J]. IET Networks. 2012, 1(3):108–115.

[10] Hadi Veisi, Hossein Sameti. Speech enhancement using hidden Markov modelsin Mel-frequency domain [J]. Speech Communication. 2013, 55(2):205–220.

[11] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, Pierre Ouellet.Front-End Factor Analysis for Speaker Veriﬁcation [J]. IEEE Trans Audio,Speech & Language Processing. 2011, 19(4):788–798.

[12] Li Deng, Xiao Li. Machine Learning Paradigms for Speech Recognition: AnOverview [J]. IEEE Trans Audio, Speech & Language Processing. 2013,21(5):1060–1089.

[13] B-H Juang. Maximum-likelihood estimation for mixture multivariate stochas-tic observations of Markov chains [J]. AT&T technical journal. 1985,64(6):1235–1249.

[14] 蒋金叶. 基于深度置信网络的文本分类方法研究 [D]. Ph.D. thesis, 2013.

[15] Geoffrey E. Hinton, Simon Osindero, Yee Whye Teh. A Fast Learning Algo-rithm for Deep Belief Nets [J]. Neural Computation. 2006, 18(7):1527–1554.

[16] 施维蒨. 门限玻尔兹曼机在人脸识别中的鲁棒性研究 [D]. Ph.D. thesis, 北京交通大学, 2015.

[17] Honglak Lee, Chaitanya Ekanadham, Andrew Y Ng. Sparse deep belief netmodel for visual area V2 [C]. Advances in neural information processing sys-tems. 2008, 873–880.

[18] Quoc V Le, Marc’Aurelio Ranzato, Rajat Monga, Matthieu Devin, Kai Chen,Greg S Corrado, Jeff Dean, Andrew Y Ng. Building high-level features usinglarge scale unsupervised learning [J]. arXiv preprint arXiv:11126209. 2011.

[19] Margarita Osadchy, Yann Le Cun, Matthew L Miller. Synergistic face detectionand pose estimation with energy-based models [J]. Journal of Machine LearningResearch. 2007, 8(May):1197–1215.

[20] Honglak Lee, Peter Pham, Yan Largman, Andrew Y Ng. Unsupervised featurelearning for audio classiﬁcation using convolutional deep belief networks [C].Advances in neural information processing systems. 2009, 1096–1104.

[21] Philippe Hamel, Douglas Eck. Learning features from music audio with deepbelief networks. [C]. ISMIR. Utrecht, The Netherlands, 2010, vol. 10, 339–344.

[22] Frank Seide, Gang Li, Dong Yu. Conversational speech transcription usingcontext-dependent deep neural networks [C]. Twelfth annual conference of theinternational speech communication association. 2011.

[23] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed,Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, BrianKingsbury, et al. Deep neural networks for acoustic modeling in speech recog-nition [J]. IEEE Signal processing magazine. 2012, 29.

[24] Fernando Rubén González-Hernández, Luis Pastor Sánchez-Fernández, SergioSuárez-Guerra, Luis Alejandro Sánchez-Pérez. Marine mammal sound classi-ﬁcation based on a parallel recognition model and octave analysis [J]. AppliedAcoustics. 2017, 119:17–28.

[25] Miguel Márquez-Molina, Luis Pastor Sánchez-Fernández, Sergio Suárez-Guerra, Luis Alejandro Sánchez-Pérez. Aircraft take-off noises classiﬁcationbased on human auditory’s matched features extraction [J]. Applied Acous-tics. 2014, 84:83–90.

[26] Luis Alejandro Sánchez-Pérez, Luis Pastor Sánchez Fernández, Sergio SuárezGuerra, José Juan Carbajal Hernández. Aircraft class identiﬁcation basedon take-off noise signal segmentation in time [J]. Expert Syst Appl. 2013,40(13):5148–5159.

[27] Selina Chu, Shrikanth S. Narayanan, C.-C. Jay Kuo. Environmental SoundRecognition With Time-Frequency Audio Features [J]. IEEE Trans Audio,Speech & Language Processing. 2009, 17(6):1142–1158.

[28] Regunathan Radhakrishnan, Ajay Divakaran, A Smaragdis. Audio analysisfor surveillance applications [C]. IEEE Workshop on Applications of SignalProcessing to Audio and Acoustics, 2005. IEEE, 2005, 158–161.

[29] Min Xu, Ling-Yu Duan, Liang-Tien Chia, Chang-sheng Xu. Audio keywordgeneration for sports video analysis [C]. Proceedings of the 12th annual ACMinternational conference on Multimedia. ACM, 2004, 758–759.

[30] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, Mingjing Li. A user attention modelfor video summarization [C]. Proceedings of the tenth ACM internationalconference on Multimedia. ACM, 2002, 533–542.

[31] STEELE Daniel, KRIJNDERS Dirkjan, GUASTAVINO Catherine. The Sen-sor City Initiative: cognitive sensors for soundscape transformations [J].

[32] María G López-Pacheco, Luis P Sánchez-Fernández, Herón Molina-Lozano. Amethod for environmental acoustic analysis improvement based on individualevaluation of common sources in urban areas [J]. Science of the Total Envi-ronment. 2014, 468:724–737.

[33] María Guadalupe López-Pacheco, Luis Pastor Sánchez-Fernández, HerónMolina-Lozano, Luis Alejandro Sánchez-Pérez. Predominant environmentalnoise classiﬁcation over sound mixing based on source-speciﬁc dictionary [J].Applied Acoustics. 2016, 112:171–180.

[34] James A Ballas. Common factors in the identiﬁcation of an assortment of briefeveryday sounds. [J]. Journal of experimental psychology: human perceptionand performance. 1993, 19(2):250.

[35] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, Mark D. Plumbley. A-coustic Scene Classiﬁcation: Classifying environments from the sounds theyproduce [J]. IEEE Signal Process Mag. 2015, 32(3):16–34.

[36] Sachin Chachada, C.-C. Jay Kuo. Environmental sound recognition: A sur-vey [C]. Asia-Paciﬁc Signal and Information Processing Association Annu-al Summit and Conference, APSIPA 2013, Kaohsiung, Taiwan, October 29 -November 1, 2013. 2013, 1–9.

[37] Oguzhan Gencoglu, Tuomas Virtanen, Heikki Huttunen. Recognition of acous-tic events using deep neural networks [C]. 22nd European Signal ProcessingConference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014. 2014,506–510.

[38] Like Xue, Feng Su. Auditory Scene Classiﬁcation with Deep Belief Network [C].MultiMedia Modeling - 21st International Conference, MMM 2015, Sydney,NSW, Australia, January 5-7, 2015, Proceedings, Part I. 2015, 348–359.

[39] 杨冬冬赵新秋, 贺海龙. 基于改进的卷积神经网络在图片分类中的应用 [C].高技术通讯. 2018.

[40] 石西建. 基于深度学习和区域图的 SAR 图像分割 [D]. Ph.D. thesis, 2014.

[41] Deep Neural Networks for Acoustic Modeling in Speech Recognition: TheShared Views of Four Research Groups [J]. IEEE Signal Process Mag. 2012,29(6):82–97.

[42] Hugo Larochelle, Yoshua Bengio, Jérôme Louradour, Pascal Lamblin. Explor-ing Strategies for Training Deep Neural Networks [J]. Journal of MachineLearning Research. 2009, 10:1–40.

[43] Yoshua Bengio, Nicolas Boulanger-Lewandowski, Razvan Pascanu. Advances inoptimizing recurrent networks [C]. IEEE International Conference on Acous-tics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada,May 26-31, 2013. 2013, 8624–8628.

[44] James Martens. Deep learning via Hessian-free optimization [C]. Proceedingsof the 27th International Conference on Machine Learning (ICML-10), June21-24, 2010, Haifa, Israel. 2010, 735–742.

[45] Nuria Oliver, Ashutosh Garg, Eric Horvitz. Layered representations for learn-ing and inferring oﬃce activity from multiple sensory channels [J]. ComputerVision and Image Understanding. 2004, 96(2):163–180.

[46] Alex Graves, Navdeep Jaitly, Abdel-rahman Mohamed. Hybrid speech recog-nition with Deep Bidirectional LSTM [C]. 2013 IEEE Workshop on AutomaticSpeech Recognition and Understanding, Olomouc, Czech Republic, December8-12, 2013. 2013, 273–278.

[47] Kunihiko Fukushima. Neocognitron: A self-organizing neural network mod-el for a mechanism of pattern recognition unaffected by shift in position [J].Biological cybernetics. 1980, 36(4):193–202.

[48] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson,Richard E. Howard, Wayne E. Hubbard, Lawrence D. Jackel. BackpropagationApplied to Handwritten Zip Code Recognition [J]. Neural Computation. 1989,1(4):541–551.

[49] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE.1998, 86(11):2278–2324.

[50] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet classiﬁca-tion with deep convolutional neural networks [J]. Commun ACM. 2017,60(6):84–90.

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh,Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bern-stein, Alexander C. Berg, Fei-Fei Li. ImageNet Large Scale Visual RecognitionChallenge [J]. International Journal of Computer Vision. 2015, 115(3):211–252.

[52] Dan C. Ciresan, Ueli Meier, Jonathan Masci, Jürgen Schmidhuber. Multi-column deep neural network for traﬃc sign classiﬁcation [J]. Neural Networks.2012, 32:333–338.

[53] Dan C. Ciresan, Alessandro Giusti, Luca Maria Gambardella, Jürgen Schmid-huber. Deep Neural Networks Segment Neuronal Membranes in Electron Mi-croscopy Images [C]. Advances in Neural Information Processing Systems 25:26th Annual Conference on Neural Information Processing Systems 2012. Pro-ceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, UnitedStates. 2012, 2852–2860.

[54] Honglak Lee, Peter T. Pham, Yan Largman, Andrew Y. Ng. Unsupervisedfeature learning for audio classiﬁcation using convolutional deep belief networks[C]. Advances in Neural Information Processing Systems 22: 23rd AnnualConference on Neural Information Processing Systems 2009. Proceedings ofa meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.2009, 1096–1104.

[55] Li Deng, Ossama Abdel-Hamid, Dong Yu. A deep convolutional neural net-work using heterogeneous pooling for trading acoustic invariance with phoneticconfusion [C]. IEEE International Conference on Acoustics, Speech and SignalProcessing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013. 2013,6669–6673.

[56] Tara N. Sainath, Abdel-rahman Mohamed, Brian Kingsbury, Bhuvana Ram-abhadran. Deep convolutional neural networks for LVCSR [C]. IEEE Interna-tional Conference on Acoustics, Speech and Signal Processing, ICASSP 2013,Vancouver, BC, Canada, May 26-31, 2013. 2013, 8614–8618.

[57] Sander Dieleman, Philemon Brakel, Benjamin Schrauwen. Audio-based MusicClassiﬁcation with a Pretrained Convolutional Network [C]. Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR2011, Miami, Florida, USA, October 24-28, 2011. 2011, 669–674.

[58] Aäron van den Oord, Sander Dieleman, Benjamin Schrauwen. Deep content-based music recommendation [C]. Advances in Neural Information ProcessingSystems 26: 27th Annual Conference on Neural Information Processing Sys-tems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe,Nevada, United States. 2013, 2643–2651.

[59] Karol J. Piczak. Environmental sound classiﬁcation with convolutional neuralnetworks [C]. 25th IEEE International Workshop on Machine Learning forSignal Processing, MLSP 2015, Boston, MA, USA, September 17-20, 2015.2015, 1–6.