ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 2,Mar. 2016,pp. 231－241
DOI:10.16337/j.1004-9037.2016.02.001

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

声学事件检测技术的发展历程与研究进展

韩纪庆

(哈尔滨工业大学计算机科学与技术学院,哈尔滨,150001)

摘　要: 声学事件检测是指对连续音频信号流中具有明确语义的片段进行检测与标定的过程。它是机器对环境声音场景进行识别和语义理解的重要基础,并将在未来类人机器人声音环境的语义理解、无人车行车周边环境的声音感知等方面发挥重要的作用。本文分别从与声学事件检测相关领域的发展历程以及应用需求出发,对声学事件检测的历史进行了回顾,介绍了典型的研究工作,并分析了未来的发展方向。在相关领域的分析中,重点介绍语音识别、基于计算的音乐处理及基于听觉特性的声音处理等方面的工作;在应用需求方面,介绍机器的环境声音感知与多媒体信息检索方面的工作;最后分析本领域的研究现状,并展望其未来的发展趋势。

关键词: 声学事件检测;语义理解;环境感知

中图分类号: TP391　　**文献标志码:** A

History and State of Art of Acoustic Event Detection

Han Jiqing

(School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China)

Abstract: Acoustic event detection refers to the task of detecting each semantic segment in an audio stream and associating it with a classification label. Acoustic event detection is a fundamental technique for sound scene recognition and semantic understanding, and it is very promising in many application fields, such as the semantic understanding of the environmental sounds for a human-like robot, the context aware of sounds in the travelling environment for an unmanned vehicle. In this paper, the history of acoustic event detection is reviewed from the point of view of related fields and application requirements, meanwhile, the typical works of acoustic event detection is introduced, and the future research of acoustic event detection is analyzed. In the analysis of related fields, we focus on the researches of speech recognition, music processing based on computation, and sound processing based on auditory. In the application requirements, we introduce the works of context aware of sounds and multimedia information retrieval. Finally, the state of the art in acoustic event detection is analyzed, and its future research fields is predicted.

Key words: acoustic event detection; semantic understanding; context aware

---

基金项目: 国家自然科学基金(91220301)资助项目。
收稿日期: 2016-01-30;**修订日期:** 2016-03-10

引 言

人类生活在一个充满声音的世界里,各种活动与事件无不伴随着丰富多彩的声音。声音是人类最为熟悉的承载信息的信号之一。人类在其漫长的发展过程中,根据积累的经验能有效地辨别出很多声音。对声音的感知与理解是人类认知这个世界的最重要途径之一。

随着信息技术的迅猛发展,开展机器模仿人类对声音认知能力的相关研究越来越受到重视,并迅速成为学术界的热点。近年来,各个国家和地区纷纷提出了与之相关的若干重大发展计划,例如欧洲的人人交互环中的计算机计划(Computers in the human interaction loop, CHIL)[1]和增强多方交互计划(Augmented multi-party interaction, AMI),美国的视频分析与内容抽取计划(Video analysis and content extraction, VACE)和学习与组织的认知助理计划(Cognitive assistant that learns and organizes, CALO)[2],以及中国国家自然科学基金委的"视听觉信息的认知计算"重大研究计划[3]等。CHIL 于 2006 年和 2007 年先后两次举办了声音感知相关的评测(Classification of events, activities and relationships, CLEAR)[4]。IEEE AASP(Audio and acoustic signal processing technical committee)也于 2013 年进行了 IEEE CASA(Computational auditory scene analysis)挑战评测[5],Barchiesi 等对 CASA 评测进行了全面总结[6]。从上述评测的结果看,声音感知与理解技术的性能还很不理想,还有较大的提升空间,仍然是一个未成熟的研究领域。

声音感知与理解的目标是使计算机能感知人耳听觉所能关注和理解的声音。声音的类型大体可分为语音和非语音。有关语音感知与理解方面已经开展了很多的研究工作。21 世纪以来,非语音感知与理解的研究已成为学术界的研究热点。非语音感知的核心技术之一就是声学事件检测(Acoustic event detection, AED),它基于数字信号处理和机器学习技术,通过采集设备拾取一段单一完整且能引起人们感知注意的短时连续声音信号,经分析处理后获得其所对应的事件表示。正如很多理论和技术的产生有其历史的必然性一样,声学事件检测技术的出现也是伴随着若干应用需求,在相关理论日渐成熟的技术支持下而产生的。本文从声学事件检测的发展历史出发,综述其相关技术的发展历程、典型工作、目前的研究现状和未来的发展趋势。

1 声学事件检测相关领域的发展历程

有关使用计算机利用数字信号处理技术来对声音信号进行处理的历史,可以追溯到 20 世纪 50 年代。伴随着电子计算机的出现和广泛应用,使用计算机来处理声音信号已经成为一种潮流,研究者们开展了大量的卓有成效的工作。由此出现了数字声音信号处理这一新的研究领域。

由于语音是人类交流中最自然的方式,因此通过机器识别人类的语音,即语音识别研究成为声音信号处理方面最早开展的工作之一。同时,音乐作为人类文化生活中重要的艺术形式,一直伴随着人们的日常生活,因此,在声音信号处理方面另一较早的工作为利用计算机对音乐的处理研究。此外,对声音的感知是人耳听觉的一个基本能力,因此从听觉特性的模拟出发来研究声音信号的处理技术也是一项重要的工作。

1.1 语音识别的发展历程

纵观语音识别的发展历程,大体经历了如下几个阶段[7,8]:20 世纪 50 年代初到 70 年代末是语音识别的起步阶段,其代表性的工作包括:(1)在 60 年代,当时苏联的 Vintsyuk 提出的采用动态规划方法来解决两个语音的时间对准问题[9],以及相关的 20 世纪 70 年代日本学者 Sakoe 提出的动态时间弯折(Dynamic time warping, DTW)算法[10],这两项工作有效地解决了语音信号的不等长处理问题;(2)有

效进行语音信号特征提取的线性预测编码(Linear predictive coding,LPC)方法[11],该阶段的研究重点是孤立词识别。20 世纪 80 年代是语音识别的发展阶段,其间出现了两项在语音识别历史上具有里程碑意义的工作:(1)基于统计的隐马尔科夫模型(Hidden Markov model,HMM)方法[12],即能对语音信号进行有效的声学建模,又能很好地与统计语言模型相结合,为有效地进行大词表连续语音识别奠定了基础。(2)将语音信号产生的声道表示模型与人耳听觉机制有效结合的特征参数表示——梅尔频率倒谱系数(Mel-frequency cepstral coefficient,MFCC)[13]。20 世纪 90 年代后,语音识别研究进入了一个在更广泛的研究领域开展工作的阶段。其代表性的工作包括:以区分性训练(Discriminative training)为代表的更精细的模型设计技术[14];以对环境噪声的影响进行补偿为代表的鲁棒识别技术[15];以采用最大似然线性回归(Maximum likelihood linear regression, MLLR)[16]、最大后验概率(Maximum a posterior, MAP)[17]准则为代表的解决训练数据不足时的各种模型自适应技术等。同时也出现了将语音识别与其他领域技术相结合的各种应用技术,如口语识别与理解、口语翻译、语音文档检索及语音情感识别等。进入 21 世纪,语音识别进入基于深度学习理论和云平台技术的全面突破阶段,识别性能显著提高[18]。2006 年加拿大多伦多大学的 Hinton 等提出了一种训练深度神经网络(Deep neural network, DNN)的方法[19]。它分为预训练和微调(Fine-tuning)两个步骤,前者利用非监督的方法逐层构建单层网络,并将其参数作为初始参数,后者通过监督的方法来获得优化后的网络参数。微软研究院最先将深度神经网络方法成功应用到语音识别中,明显降低了误识率,从而引发了基于深度神经网络的语音识别热潮[20]。目前无论是学术机构、还是工业界都投入大量的人力和财力致力于此方面的研究。

1.2 基于计算的音乐处理发展历程

从基于计算的音乐处理的发展历程看,最早的工作可以追溯到 1928 年,由前苏联科学家列昂泰勒明(Leon Theremin)发明的泰勒明电子琴[21]。它将一根天线连接到一个带有放大器和扬声器的振荡电路上,利用天线和演奏者的手构成电容器,通过天线接收手的位置变化来发出声响。泰勒明电子琴是目前为止唯一不需要身体接触的电子乐器。其后,泰勒明电子琴被广泛应用于 20 世纪 40 至 50 年代好莱坞电影配乐中[22,23]。同样在 20 世纪 40 年代,伴随着示波器的出现,以及数字信号处理技术的广泛应用,引发了电子音乐发展的第 1 次浪潮。利用示波器,研究者可以在阴极射线管上观测波形的稳定部分,并通过计算傅里叶级数来进行音乐的分析。典型的工作是分析管乐器的音调,以确定其是否与弦乐器一样具有共振性。同时,很多著名的作曲家都使用从电子实验室中获得的信号发生器来创作音乐[24]。

20 世纪 60 年代出现了通用数字计算机,不久即被用于音乐的分析与合成中。最早使用计算机进行乐器音调分析的一项工作是由美国麻省理工学院的 Luce 完成。他分析了大量不同乐器的音调,以深入了解这些乐器的工作状态,这一工作也成为后续很多工作的基础[25]。另一重要的工作来自于美国伊利诺伊大学的 Freedman,他将乐器的音调用一组分段常数频率的正弦函数和的形式来建模,正弦函数的振幅为指数与常数的分段和,使用线性插值来平滑频率之间的变化[26]。

20 世纪 70 年代,伴随着语音信号处理技术的发展,很多语音信号处理中的特征提取方法,尤其是基音检测等技术被借鉴应用到了音乐信号处理中,并且获得较好的性能。美国斯坦福大学的 Moorer 研究了音乐的转写(Transcription)技术,通过对一段复调的音乐声音进行数字化处理,使其转化为所演奏音符对应的文字形式描述的音乐符号[24]。Moorer 的方法首先使用一个基音检测器确定每个时间片段上的和声,并将其用于确定一个带通滤波器,以保证乐器的每个谐波都至少通过一个滤波器。因此,每个滤波器的输出都经过了基音检测和能量检测的处理。通过上述的处理能给出作为时间函数的能量和频率信息,每对能量和频率函数作为其特征。然后利用多组能量和频率函数对来推断出音符,其中能量

和函数对都是同步出现在与谐波有关的频率上。接着通过将音符分离为高、低音等若干组进而获得旋律,之后进行音乐的转写。

尽管在语音信号处理研究中积累的很多技术可以移植到音乐信号处理中,但音乐信号本身也拥有其自身的声学特征和结构特征,使其有别于口语或其他的非音乐信号。为此,从 20 世纪 80 年代起,研究者们开展了很多针对音乐本身特点的技术研究。典型的如美国斯坦福大学开展了针对不同乐器和音乐风格的复调音乐的识别研究,通过使用声学知识、上下文信息及源相干(Source coherence)信息来改进性能。这些工作能为不同的音乐表演、音乐转写以及数字音频文件的分割提供工具[27]。

20 世纪 90 年代后,音乐信号处理的研究范围日益广泛。典型的包括:(1)音符起始点(Note onset)检测研究。它是指确定音符或其他音乐事件开始的物理位置。这一工作是很多音乐分析的基本步骤,以及音乐检索时索引构建的基础。音符起始点检测的通常方法是寻找信号中短暂的区域,例如突然的能量爆发、信号短时频谱或统计特性上的变化等。通常情况下,音符的起始点与上述的短暂区域的起始点一致[28]。(2)音乐体裁分类。音乐体裁是一个关键的描述,它是由音乐创作人和图书馆管理员在组织音乐作品集时使用的高层描述。其广泛用于音乐目录的组织、图书馆及音乐商店中。这种将一段音乐与某一体裁相关联的方式有助于帮助用户找到其所要的音乐。尽管人们广泛使用诸如爵士、摇滚和流行等概念来说明音乐的体裁,但音乐的体裁在定义上仍遗留着相应的问题,很多音乐并非能用简单的文字来充分描述,这使得对其自动分类是一个很难处理的问题。尽管如此,研究者们仍开展了卓有成效的工作,从旋律、和声、节奏、音色以及空间位置等角度提取音乐的特征,并从 k 近邻、高斯混合模型、隐马尔科夫模型、支持向量机及人工神经网络等来构建分类器[29]。

1.3 基于听觉特性的声音处理发展历程

在语音识别的早期研究中,对语音特征的提取主要基于人类的发音机理,通过工程化的方法来模拟发音时的声道特性,经典的特征如共振峰特征、线性预测系数特征等,很少涉及人耳听觉机理的特性。这一方面源于声道模型中的主要器官口腔相对开放,较易进行直观的分析,模型构造起来相对简单;另一方面相对于人类的发音机理,人类的听觉机理非常复杂,很难进行非破坏性的物理观测,因此对听觉机理的研究更多是心理学的主观评测研究,而且即使在听觉机理中已经取得了相应成果,对信号处理领域的研究者来说,要么并没有关注这些进展,要么不知道如何加以工程实施。例如在语音处理中最成功的特征之一 MFCC 特征[13],其实仅模拟了耳蜗对高低频不同频带的敏感程度,即人耳对低频信号比对高频信号更敏感,进而将频率轴按 Mel 频率刻度进行不均匀划分。而听觉认知中有关 Mel 频率划分的研究成果最早出现在 20 世纪 30 年代,Stevens 等人的经典文献[30],但直到 20 世纪 80 年代才出现了 MFCC 特征。MFCC 已经将人耳听觉认知的特性反映在语音信号处理的特征提取中,并且验证了其性能明显优于先前的线性预测倒谱系数。但 MFCC 仅仅反映了耳蜗对不同声音频带的滤波作用,如何将更多的听觉认知特性引入到语音信号处理的研究中,以期获得更好的语音识别性能,并没有取得很大的进展。

从 20 世纪 80 年代中期开始,语音识别面临的一个主要困难就是噪声处理。很多在实验室安静环境下表现优异的系统,当应用到实际有噪声的环境时,性能显著下降。从那个年代开始,实际噪声环境下鲁棒的语音识别研究一直是本领域一个重要的研究方向。研究者们从特征提取和模型构建两个方面开展了研究。从特征提取方面看,单纯地在基于发音机理的声道模型方面寻找突破已经很困难。这促使人们将希望的目光投向了听觉机理的工程化处理上。研究者们逐渐意识到有效地进行声音处理的关键问题之一是要有一个合适的人耳听觉模型。因而开始更多地关注于听觉心理学与听觉生理学上的研究进展。

20 世纪 70 年代以前,听觉心理学在人耳的低层次检测分离研究上取得了较大的成功,例如耳间时间差、耳间声级差以及耳蜗的滤波特性等,但听觉上的研究进展不如视觉那么大[31]。20 世纪 70 年代以后,听觉研究转向了更深层次的工作,例如双耳效应、听觉的空间定位与跟踪以及多信息流的分离等[31]。加拿大著名听觉心理学家布雷格曼 1990 年出版的经典学术专著《听觉场景分析》(Auditory scene analysis, ASA)[32]对声音处理起到重要指导意义的听觉认知机理的里程碑式的研究工作。布雷格曼借鉴了计算机视觉处理中场景分析的概念,使用听觉场景分析来命名人耳对声音的分辨能力。其定义的听觉场景分析,主要是指人耳具有能够在复杂的声学环境中区分出各个独立声源的能力。在这部专著中,他给出了人耳听觉系统对混合声音的检测和分离的一系列准则,从而为包括声学事件检测在内的很多研究提供了听觉感知处理上的理论依据。听觉场景分析不仅获得了心理学家们的关注,也为信息处理工作者利用计算机通过构建模型来进行声音处理提供了理论指导,由此出现了计算听觉场景分析(Computational auditory scene analysis, CASA)[33]。

计算听觉场景分析理论为语音识别等领域更好地将人耳听觉认知的机理应用其中提供了重要的依据。事实上,随着在实际噪声环境下语音识别鲁棒性研究的深入,对公共场所等多人混杂时的噪声处理问题日显突出。没有特殊处理的语音识别系统很难区分出哪个语音信号是其应关注的信号。人耳具有一种先天的听力选择能力,即在复杂的声学环境中能将注意力集中在某个人的谈话之中,而忽略背景中其他的对话或噪音,这就是所谓的鸡尾酒会效应(Cocktail party effect)[34]。计算听觉场景分析为解决语音识别中的鸡尾酒会效应问题提供了手段。

20 世纪 90 年代,将听觉心理学上的研究成果应用到语音信号处理中的另一项重要工作是 Hermansky 提出的感知线性预测(Perceptual linear predictive, PLP)技术[35]。他将人耳听觉试验获得的一些结论,通过近似计算的方法进行工程化的处理,再用简化的模型进行模拟后应用到频谱分析中。经过这样处理后的语音频谱考虑到了人耳的听觉特点,因而有利于语音信号处理。一些研究表明,对噪声环境下的语音识别,采用感知线性预测特征比美尔倒谱系数特征的性能更好。尽管感知线性预测已经对听觉的各种特性进行了相应的简化,但其各个计算步骤还是相当复杂,运算量较大。

从听觉生理学研究的领域看,其主要开展耳蜗和听觉中枢在听觉认知中的机理与作用的研究。耳蜗由内毛细胞(Inner hair cell)、外毛细胞(Outer hair cell)、基底膜(Basilar membrane)和覆膜(Tectorial membrane)等构成。它能将接收到的声音信号通过复杂且高效的处理机制转换为神经信号,并传递给听觉中枢。听觉中枢能对接收到的听神经信号先进行表示与编码,接着进行相关分析,将相应成分整合为声模式,之后对声音信号进行辨识[36]。

20 世纪 70 年代开始,有关耳蜗机理的研究取得了显著性的进展。Russell 等在研究内毛细胞的交流和直流电位时发现,其交流成分的频率与刺激声完全一致,与从外耳道内记录的耳蜗微电位的波形非常吻合[37]。Dallos 在研究外毛细胞内电位时发现,其结果与内毛细胞的交流电位非常相似[37]。在基底膜的功能及作用研究方面,由于基底膜的不同位置对声音的响应过程类似于一个滤波的过程,为此研究者们尝试设计了多种滤波器来模拟基底膜的滤波特性。其中应用最广泛的是 GammaTone 滤波器。GammaTone 函数最早是由 Johannesma 引入到听觉研究中[38]。从时域波形看,它是一个振动频率等于其中心频率,振动包络为 Gamma 函数曲线的波形。Boer 等进一步使用该函数来模拟基底膜的滤波特性,并将其命名为 GammaTone 滤波器[39]。此后,Patterson 等给出了后来广泛使用的 GammaTone 滤波器的形式[40]。

进入 21 世纪后,关于基底膜和覆膜间耦合机制的研究取得了明显进展。Russell[41]和 Ghaffari[42]等的研究发现,基底膜和覆膜间的耦合机制,表现出尖锐的频率选择性增益特性,它是听觉产生的基础,并

对听觉的区分性和鲁棒性具有非常重要的影响[36]。

近年来,在听觉中枢方面的研究工作取得了重要进展[36]。Smith 等的研究发现,听觉中枢在表示和编码声音信号时,是以最小的能量消耗与神经数量编码并传递最大化信息的方式,来对声音信号进行编码[43]。Hill 等的研究表明,听觉中枢以声模式的方式对声音信号进行分析和识别,且声音信号中具有统计独立的内蕴时频结构对单信道情况下听觉声源分离具有重要影响[44]。

基于上述听觉生理学的研究进展,研究者们提出了很多基于听觉机理的声音信号的特征提取方法。尤其是近年来,随着听觉中枢对声音信号编码方式的发现,以及数字信号处理领域稀疏表示与压缩感知理论的提出,出现了若干采用稀疏表示来近似听觉中枢编码方式的特征提取方法[45-48]。

2 声学事件检测的应用需求

声学事件检测的应用需求主要来源于两个方面:(1)机器的环境声音感知;(2)基于语义的多媒体信息检索。两者尽管面向不同的应用领域,但无一例外地需要有效的声学事件检测技术。

2.1 机器的环境声音感知

20 世纪 90 年代,从计算模式的发展趋势看,普适计算逐渐兴起。人们认识到未来的计算机将无处不在,可能集计算机、手机、传呼机及收音机等于一体。计算的物理位置正从传统的桌面方式逐步向以嵌入式处理为特征的无处不在的方式发展。从应用需求看,随着人们可以获得的各种各样的大量信息源数据的增加,使用电子邮箱、声音邮箱(Voice mail)进行通讯的人数日益增多。人们越来越希望在诸如会议中、课堂上、驾驶或走路等远离桌面计算机的情况下,仍能无缝地(Seamless)获得个人信息和通讯服务。因此,无论从计算模式的转变,还是从实际应用的牵引看,都迫切需要研究诸如手持计算机、可穿戴(Wearable)计算机等具有移动计算能力的设备。

为了提高移动计算设备的智能化,研究人性化的人机交互输出技术十分必要。采用人性化输出技术的系统,能够根据用户所处的时间、场景及信息的重要程度等情况,自主地选择最需要输出的信息类型和信息内容[49]。在信息化社会中信息源很多,用户可能感兴趣的信息也很多,系统应根据相关的条件来筛选和安排信息的输出顺序,同时为了保证不使有用的信息丢失,需要对相应的还没有输出的信息进行保存,当用户需要这些信息时,能将其以语音合成的方式进行输出[49]。为了帮助用户记忆正在播放的信息类型或到达时间等,可以将不同信息类型或到达时间等映射到不同的听觉感知的空间位置上,这样通过在不同感知空间位置上播放信息,能让用户了解信息的类型和到达时间等情况[49]。在此方面美国麻省理工学院媒体实验室较早开展了比较全面的工作,其所开发的演示系统 Nomadic Radio[50],将音频输入输出集中到一个可穿戴平台上[51],采用语音作为信息输入方式,能根据时间、场景[52]等提供给用户整点新闻、电子邮箱、天气预报和股票等信息。麻省理工学院将这种技术称为可穿戴音频计算(Wearable audio computing)[53]技术。

随着移动和可穿戴计算设备研究的兴起,人们注意到使用移动设备的用户往往在不同的环境和状态下从事不同的活动,因此希望其随身携带的移动设备能自动判断出所处的环境,并作出相应的调整。由此出现了环境感知(Context aware)技术的研究。环境感知这一概念最初来自于普适计算[54],其目的是使用计算机通过处理环境中相关联变化的情况,进而能感知其所处的环境并作出反应。使用这种环境感知的系统也能侧面反映出其使用者所处的环境状态。环境感知系统的处理流程通常包含如下几个步骤:使用传感器来采集环境信息、对环境信息进行抽象和理解以及基于识别结果进行行为决策。由于在很多应用中用户的活动和位置非常关键,因此,早期的环境感知工作更多地关注于位置感知和行为识别方面的研究。后续的研究中也涉及了与所处环境相关的诸如视听觉等物理特性的感知研究。

实现环境感知的关键之一就是机器能够感知和判定其所处的环境,包括能识别出环境中图像和声音,甚至触摸的方式,以便更好地实现人机交互。机器感知技术不仅能处理用户的直接命令,而且能充分利用周边的音视频输入以预测出可能对其有用的额外信息。具体对声音环境而言,即需要机器的环境声音感知技术。机器的环境声音感知是通过采集环境中的声音信号,通过相关的分析技术来判定所处的环境类型。一个典型的例子为能连续感知其周边环境的智能手机。使用这种手机,用户希望通过手机上的环境声音感知技术能够判断出其当前是在会议室中,还是在嘈杂的公路上;这样当在会议室时,可将手机铃声自动切换到振动或静音状态,而在公路上时,将手机铃声尽可能调大以便当有来电时能及时听到。同时,具有上述功能的手机可以识别出其用户是在会议室中,并判断出用户是否正在开会,进而拒绝一些不重要的来电[55]。

美国麻省理工学院媒体实验室的 Sawhney 和 Maes 完成最早的机器环境声音感知的工作,他们首次提出了相应的方法[56]。在研究中录制了来自包含人群、地铁和交通等环境类型的数据集,通过从这些数据中提取特征,并采用回归神经网络和 k 近邻方法进行分类,获得了 68%的分类精度。其后,还是上述实验室的研究人员通过使用一辆自行车骑行到超市,利用随身携带的穿戴设备上的麦克风录制了整个行程中,包括家里、街道和超市的连续音频流,之后将这一音频流自动切分为不同的音频场景,提取各场景特征后使用 HMM 进行分类[57],获得了初步的识别结果。实际上,这一研究工作与前面孤立的声音环境的感知不同,已经开启了连续声音流中不同声音环境感知的研究工作。进入 21 世纪,人们越来越关注于将听觉认知特性应用到声音的环境感知中。鉴于心理声学中强调局部和全局特性在环境声音感知中均有作用,研究者开始在环境声音感知中同时考虑使用局部和全局特性。经典的工作是由 Eronen 等[58]完成,他们采用 MFCC 特征来描述音频信号的局部频谱包络,基于高斯混合模型(Gaussian mixture model,GMM)描述它们的统计分布。之后采用一个能利用训练数据类别知识的区分性算法来训练 HMM,以反映 GMM 在时间上的变化。进一步地,他们还通过考虑多种特征,以及在分类算法中加入特征变换步骤来改进性能,对 18 类不同声学场景获得了 58%的分类精度[59]。

机器的环境声音感知作为听觉认知研究的一个重要组成部分,不仅引起了欧美等发达国家的重视,同样也引起了中国政府的重视。2008 年中国国家自然科学基金委紧跟国际学术前沿,适时启动了"视听觉信息的认知计算"重大研究计划,旨在从人类的视听觉认知机理出发,研究并构建新的计算模型与计算方法[3],重点探讨解决"感知特征提取、表达与整合"、"感知数据的机器学习与理解"和"多模态信息协同计算"等核心科学问题。通过上述问题的解决,拟构建智能车辆无人驾驶验证平台[3]。

无人驾驶车辆是一种集感知、控制和智能决策等理论与技术于一体,能够自主驾驶的智能车辆。它充分体现了信息技术、控制技术和计算机技术等诸多领域的综合实力,无论在军事方面,还是在民用方面都有着广泛的应用前景[60]。无人车研究的核心内容之一是智能行为决策,而智能行为决策的前提则是其行驶过程中对周边环境的自动感知。无人车感知环境信息的手段可以有多种,例如全球定位系统、激光、雷达、红外线及视听觉信息等。其中视听觉信息的自动感知在无人车的行驶中占有重要的地位。目前使用较多的是视觉感知信息,而较少利用听觉感知信息。事实上,听觉信息也是无人车系统不可或缺的重要决策依据信息:一方面听觉感知结果是对视觉感知结果的重要补充,通过两者的有效结合可以更准确地感知车辆的周边环境,这种辅助作用在诸如黑夜或隧道等可利用的视觉信息不理想的情况下,就表现得尤为突出;另一方面外部世界与无人车间的很多交互信息基于声音,例如警车和救护车的警笛声、铁路道口的警示声及各种车辆提示避让的鸣笛声等。感知周围这些基于声音的交互信息,并做出正确的智能决策对无人车而言至关重要[60]。因此,环境声音的感知也能为无人车提供重要的辅助信息。

哈尔滨工业大学联合中国科学院自动化所承担了"视听觉信息的认知计算"重大研究计划中有关无

人车行车环境听觉模型及声音处理关键技术的项目,重点开展行驶中的无人车辆对车内外声音的自动检测、实时识别和理解方面的研究。为实现机器的环境声音感知,其关键的工作即是检测环境中的声学事件,因此对机器环境声音感知的迫切需求,推动和促进了声学事件检测的研究。

2.2 多媒体信息检索

从多媒体信息检索方面看,随着现代信息技术,特别是多媒体与网络技术的发展,多媒体数据库中的内容越来越多,亟需有效地识别、标注和检索等技术。相比于音视频等多媒体数据的检索,文本数据的检索发展最早且进步最快。因此,早期对多媒体数据的检索,主要是基于文本的检索。它首先通过人工来对多媒体数据进行识别分类,然后为其标注文字标签,检索时则通过文本方法进行。这种方法有其固有的缺陷:人工根本无法完成数量庞大且呈指数增长的海量数据的识别和标注任务;同时也无法实现灵活的个性化识别和检索任务,由此出现了基于音视频多媒体内容的检索研究。

在基于内容的多媒体信息检索中,音频内容检索占有重要的地位。这主要有两方面的原因:(1)现实中存在着大量的音频数据,例如各种音乐、音效、语音文档、广播节目及会议录音等。采用音频信息检索技术,能够从大量的音频数据中快速准确地查找到用户感兴趣的音频信息。例如用户需要某个具有特殊主题场景的音频片段,例如婚礼、生日及聚会等场景音频。这类个性化的检索任务无法由简单的文字检索来实现。(2)音频信息检索可以作为辅助手段来实现音视频多媒体信息的检索。例如在电影、电视剧的片头等处,视频信息通常变化剧烈,而片头曲等音频信息却保持稳定,始终表示同一语义。如果只是按照视觉特征对其分类,则这些多媒体数据流就可能会被分成为不同的语义场景;而如果按音频特征进行分类,就可以将它们划分到同一语义场景中。

鉴于音频场景大都是由多种声学事件组成,通过检测音频信号中相应的声学事件,将它们作为特定的语义加以利用,无论是对音频信息检索中的索引构建,还是对搜索各个声学事件所反映的特定语义片段都十分重要。因此,有效的声学事件检测是检索特定音频场景的关键技术之一。由此看来,为高效实现基于内容的多媒体信息检索也是催生声学事件检测研究的另一助推器。

3 声学事件检测的现状及发展趋势

从声学事件检测的发展脉络可以看出:多年来在语音识别和音乐处理方面的研究工作,为声学事件检测提供了数字信号处理与机器学习层面的技术积累。听觉认知机理研究方面的工作使声学事件的检测能更多地模拟人耳的听觉功能,以获得更好的检测性能。机器的环境感知以及基于语义的多媒体信息检索,对声学事件的检测提出了强烈的需求。正是在上述多种因素的驱动下,声学事件检测的研究获得了飞速发展。

从总体上看,声学事件检测的研究经历了从简单事件类型到复杂事件类型的检测;从孤立片段的事件检测到连续声音流中的事件检测;从实验室模拟的声学事件到现实生活中的声学事件检测的过程。早期的工作一般只针对特定声音,同时只涉及到较少的声音类型,且不同声音间很少有交叠。此外,仅在较小的数据集上进行测试。因此,早期的方法大多不能应用于现实生活中连续音频流的声学事件检测。目前的研究更多关注于真实环境下连续音频流中的声学事件检测研究。近年来,数字信号处理与机器学习中,诸如稀疏表示与压缩感知、深度学习等方面的长足发展,为声学事件检测研究提供了更有效的理论方法和技术手段,促进了声学事件检测性能的持续提高。目前真实环境下声学事件检测的性能还很不理想,在未来的研究中,除了不断地利用更有效的机器学习方法外,更多地关注听觉认知机理的最新进展,将其研究成果应用于声学事件检测之中是一项非常有意义的工作。同时,目前较好的检测方法计算复杂度都偏高,如何减低计算复杂度也是非常重要的工作。此外,开展基于多个声学事件检测

结果的声音段落的整体语义理解也是更重要的研究工作。

4 结束语

智能化、人性化的人机交互是近年来学术界的研究热点之一。为实现此目标，迫切需要开展机器自动感知周边环境的研究，其主要工作之一就是环境声音感知的研究。声学事件检测是环境声音感知研究的基础性工作。因此，开展声学事件检测的研究具有重要的理论意义和实用价值。本文对声学事件检测的发展历程进行了回顾，介绍了典型的相关工作，并就其未来的发展趋势进行了展望。从中可以看出，声学事件检测是一个极具潜力的研究课题，还有许多问题需要解决。

参考文献：

[1] Waibel A, Stiefelhagen R, Carlson R, et al. Computers in the human interaction loop, handbook of ambient intelligence and smart environments[M]. NY, USA: Springer, 2009:1071-1116.

[2] Stiefelhagen R, Bernardin K, Bowers R. Multimodal technologies for perception of humans[M]. Berlin, Garman: Springer-Verlag Berlin Heidelberg, 2008.

[3] 国家自然科学基金委员会. “视听觉信息的认知计算”重大研究计划[EB/OL]. http://ccvai.xjtu.edu.cn/, 2008-08-26.
National Natural Science Foundation of China. Major research plan of "Cognitive computing of visual and auditory information"[EB/OL]. http://ccvai.xjtu.edu.cn/2008-08-26.

[4] Stiefelhagen R, Bernardin K, Bowers R, et al. Classification of events, activities and relationships[C]// International Evaluation Workshops CLEAR 2007 and RT 2007. Baltimore, USA: Springer, 2007:3-34.

[5] Stowell D, Giannoulis D, Benetos E, et al. Detection and classification of audio scenes and events[J]. IEEE Trans on Multimedia, 2015,17(10):1733-1746.

[6] Barchiesi D, Giannoulis D, Stowell D, et al. Acoustic scene classification: Classifying environments from sounds they produce [J]. IEEE Signal Processing Magazine, 2015,5:16-34.

[7] O'Shaughnessy D. Automatic speech recognition: History, methods and challenges[J]. Pattern Recognition, 2008,41:2965-2979.

[8] 倪崇嘉，刘文举，徐波. 汉语大词汇量连续语音识别系统研究进展[J]. 中文信息学报，2009,23(1):112-123.
Ni Chongjia, Liu Wenju, Xu Bo. Research on large vocabulary continuous speech recognition systems for mandarin Chinese [J]. Journal of Chinese Information Processing, 2009,23(1):112-123.

[9] Vintsyuk T. Speech discrimination by dynamic programming[J]. Kibernetika, 1968,4(1):81-88.

[10] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE Trans on ASSP, 1978,26:43-49.

[11] Saito S, Itakura F. The theoretical consideration of statistically optimum methods for speech spectral density[R]. Report No: 3107, Tokyo, Japan: NTT, 1966.

[12] Baker J. The dragon syatem—An overview[J]. IEEE Trans Acoust Speech and Signal Processing, 1975,23(1):24-29.

[13] Davis S, Mermelstein P. Comparison of parametic representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Trans on Acoustics, Speech, and Signal Processing, 1980,28(4):357-366.

[14] Juang B H, Katagiri S. Discriminative learning for minimum error classification[J]. IEEE Trans on Signal Processing, 1992, 40(12):3043-3054.

[15] Junqua J C, Haton J P. Robustness in automatic speech recognition: Fundamentals and applications[M]. Boston: Kluwer Academic Publishers, 1996.

[16] Leggetter C J, Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models[J]. Computer Speech and Language, 1995,9(2):171-185.

[17] Gauvain J L, Lee C H. Maximum a posterior estimation for multivariate Gaussian observation[J]. IEEE Trans on Speech and Audio Processing, 1994,2(2):291-198.

[18] 戴礼荣，张仕良. 深度语音信号与信息处理：研究进展与展望[J]. 数据采集与处理，2014,29(2):171-179.
Dai Lirong, Zhang Shiliang. Deep speech signal and information processing: Research progress and prospect[J]. Journal of Data Acquisition and Processing, 2014,29(2):171-179.

[19] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006,313(5786): 504-507.

[20] Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks[C]// Annual Conference of International Speech Communication Association (Interspeech2011). Florence, Italy: International Speech Communication Association, 2011:437-440.

[21] Glinsky A. Theremin: Ether music and espionage [D]. Urbana, Illinois: University of Illinois Press, 2000.

[22] Roads C. The computer music tutorial[M]. Cambridge, MA, USA: MIT Press, 1996.

[23] Muller M, Ellis D P W, Klapuri A, et al. Signal processing for music analysis[J]. IEEE Journal of Selected Topics in Signal Processing, 2011,5(6):1088-1109.

[24] Moorer J A. On the segmentation and analysis of continuous musical sound by digital computer[D]. Stanford, USA: Stanford Univerisity, 1975.

[25] Luce D A. Physical correlates of nonpercussive musical instrument tones[D]. Cambridge: MIT, 1963.

[26] Freedman M D. A technique for analysis of musical instrument tones[D]. Urbana: USA, University of Illinois, 1965.

[27] Chafe C, Jaffe D. Source separation and note identification in polyphonic music[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP86). Tokyom, Japan: IEEE Signal Processing Society Press, 1986:1289-1992.

[28] Bello J P, Daudet L, Abdallah S, et al. A tutorial on onset detection in music signals[J]. IEEE Trans on Speech and Audio Processing, 2005,13(5):1035-1047.

[29] Scaringella N, Zoia G, Mlynek D. Automatic genre classification of music content: A survey[J]. IEEE Signal Processing Magazine, 2006,3:133-141.

[30] Stevens S S, Volkmann J, Newman E B. A scale for the measurement of the psychological magnitude pitch[J]. Journal of the Acoustical Society of America, 1937,8(3):185-190.

[31] 吴镇扬,张子瑜,李想,等.听觉场景分析的研究进展[J].电路与系统学报,2001,6(2):68-73.
Wu Zhenyang, Zhang Ziyu, Li Xiang, et al. The research advance of auditory scene analysis[J]. Journal of Circuits and Systems, 2001,6(2):68-73.

[32] Bregman A S. Auditory scene analysis[M]. Massachusetts, USA: MIT Press, 1990.

[33] Wang D L, Brown G J. Computational auditory scene analysis: Principles, algorithms and applications [M]. USA: IEEE Press/Wiley-Interscience, 2006.

[34] Bronkhorst, Adelbert W. The cocktail party phenomenon: A review on speech intelligibility in multiple-talker conditions[J]. Acta Acustica United with Acustica, 2000,86:117-128.

[35] Hermansky H. Perceptual linear predictive (PLP) analysis of speech[J]. Journal of Acoustical Society of America, 1990,87(4):1738-1752.

[36] 游大涛.基于听觉机理的鲁棒特征提取及在说话人识别中的应用[D].哈尔滨:哈尔滨工业大学,2013.
You Datao. Auditory mechanism based robust feature extraction and its application in speaker recognition[D]. Harbin: Harbin Institute of Technology, 2013.

[37] 赵立东,曹效平,贾学斌,等.听觉生理学研究的进展及未来(一)[J].中华耳科学杂志,2013,11(3):353-356.
Zhao Lidong, Cao Xiaoping, Jia Xuebin, et al. Research progress and prospect of auditory physiology [J]. Chinese Journal of Otology, 2013,11(3):353-356.

[38] Johannesma P I M. The pre-response stimulus ensemble of neurons in the cochlear nucleus [C]// Symposium on Hearing Theory. Holland: BL Cardozo, 1972:58-69.

[39] De Boer E, De Jongh H. On cochlear encoding: Potentialities and limitations of the reverse-correlation technique[J]. The Journal of the Acoustical Society of America, 1978,63:115-135.

[40] Patterson R, Robinson K. Holdsworth J, et al. Complex sounds and auditory images[C]// Proceedings 9th International Symposium on Hearing. France: Pergamon Press, 1992:429-446.

[41] Russell I, Legan P, Lukashkina V, et al. Sharpened cochlear tuning in a mouse with a genetically modified tectorial membrane[J]. Nature Neuroscience, 2007,10(2):215-223.

[42] Ghaffari R, Aranyosi A, Freeman D. Longitudinally propagating traveling waves of the mammalian tectorial membrane[J]. Proceedings of the National Academy of Sciences, 2007,104(42):16510-16515.

[43] Smith E, Lewicki M. Efficient auditory coding[J]. Nature, 2006,439:978-982.

[44] Hill K, Miller L. Auditory attentional control and selection during cocktail party listening[J]. Cerebral Cortex, 2010,20(3):583-590.

[45] Ness S, Walters T, Lyon R. Auditory sparse coding[J]. Music Data Mining, 2012,21:77-97.

[46] Lyon R F, Ponte J, Chechik G. Sparse coding of auditory features for machine hearing in interference[C]// IEEE Interna-

tional Conference on Acoustics, Speech and Signal Processing (ICASSP2011). Prague, Czech Republic: IEEE Signal Processing Society Press, 2011,2:5876-5879.

[47] You D, Jiang T, Han J, et al. A cochlear Neuron based robust feature for speaker recognition[C]//Proc International Conference on Acoustics, Speech, and Signal Processing (ICASSP2011). Czech Republic: IEEE Signal Processing Society Press, 2011:5440-5443.

[48] 孙林慧,杨震.语音压缩感知研究进展与展望[J].数据采集与处理,2015,30(2):275-288.
Sun Linhui, Yang Zhen. Compressed speech sensing for research progress and prospect[J]. Journal of Data Acquisition and Processing, 2015,30(2):275-288.

[49] 韩纪庆,张磊,吕成国,等.可穿戴计算机中的语音处理技术[J].计算机科学,2002,29(5):107-109.
Han Jiqing, Zhang Lei, Lü Chengguo, et al. Techniques of speech processing for wearable computer[J]. Computer Science, 2002,29(5):107-109.

[50] Sawhney N, Schmandt C. Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments [J]. ACM Transactions on Computer Human Interaction, 2000,7(3):353-383.

[51] Sawhney N, Schmandt C. Speaking and listening on the run: Design for wearable audio computing[C]//Proceedings of the International Symposium on Wearable Computing. Pennsylvania, USA: IEEE Signal Processing Society Press, 1998:108-115.

[52] Clarkson B, Pentland A. Extracting context from environmental audio[C]//Proceedings of the International Symposium on Wearable Computing. Pennsylvania, USA: IEEE Signal Processing Society Press, 1998:154-155.

[53] Roy D, Sawhney N, Schmandt C, et al. Wearable audio computing: A survey of interaction techniques[R]. MIT Media Lab Technical Report, Cambridge, USA: MIT, 1997:1-9.

[54] Schilit B, Adams N, Want R. Context-aware computing applications[C]//IEEE Workshop on Mobile Computing Systems and Applications. Santa Cruz, USA: IEEE Signal Processing Society Press, 1994:89-101.

[55] Schmidt A, Aidoo K A, Takaluoma A, et al. Advanced interaction in context[C]//1st International Symposium on Handheld and Ubiquitous Computing(HUC99). Germany: Springer LNCS, 1999, 1707:89-101.

[56] Sawhney N, Maes P. Situational awareness from environmental sounds[R]. Techn Rep Massachusetts Institute of Technology, Cambridge, USA: MIT, 1997:1-7.

[57] Clarkson B, Sawhney N, Pentland A. Auditory context awareness via wearable computing[C]//Proc 1998 Workshop Perceptual User Interfaces(PUI'98). San Francisco, USA: ACM, 1998:1-9.

[58] Eronren A J, Peltonen V T, Tuomi J T, et al. Audio-based context awareness-acoustic modeling and perceptual evaluation [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2003). Hong Kong, China: IEEE Signal Processing Society Press, 2003:529-532.

[59] Eronren A J, Tuomi J T, Klapuri A, et al. Audio-based context recognition[J]. IEEE Trans on Audio, Speech, Language Processing, 2006,14(1):321-329.

[60] 孔鸿运. 行车环境下鲁棒的声学事件检测方法[D]. 哈尔滨:哈尔滨工业大学,2013.
Robust acoustic event detection methods in driving environment[D]. Harbin: Harbin Institute of Technology, 2013.

作者简介:

韩纪庆(1964-),男,教授,博士生导师,研究方向:语音信号处理、音频信息处理等,E-mail:jqhan@hit.edu.cn。