1

第 1 单元

人工智能的概念

1.1 智能机器的梦想

学习目标

（1）理解从古至今人类对智能机器的梦想与追求。

（2）了解历史上具有代表性的自动化机器及其影响。

人们很早就希望制造出聪明的机器，能够与人聊天或者帮助人们做事。关于智能机器的梦想可以追溯到很早以前，但这些大多只是传说，并没有实际证据表明确有其事。后来，随着技术的进步，出现了一些自动化的机械装置，部分满足了人们关于智能机器的梦想。而真正智能的机器是在计算机诞生以后才出现的。这些机器用计算的方式来模拟人类的思维，最终实现了人类的千年梦想。

① 古代智能机器的传说

古人对智能机器怀有强烈的向往。例如，在《墨子·鲁问》中记载了鲁班削竹为鹊（图1-1）的故事，说的是一位叫鲁班的巧匠，他用竹子制作了一个飞鸟，可以在天上连续飞好几天。“削竹木以为鹊，成而飞之，三日不下”。后人猜测鲁班制作的很可能是个风筝，只是由于技艺高超，使得风筝能在天上飞的时间很长。也有人戏称鲁班是第一台无人机的发明者。据说与鲁班同时代的墨子也擅长制作各种灵巧的机器，他也造了一只会飞的鸟，只不过看起来比鲁班的鸟要差一些，飞了一天就掉下来了。

图1-1　鲁班削竹为鹊示意

在《列子·汤问》中还记载了一位与鲁班同时代的巧匠，名叫偃师。偃师比鲁班和墨子还要厉害，可以用木头、毛发、油漆等材料制造出惟妙惟肖的人偶，从外表看和真人无异，而且能歌善舞。偃师带着这个人偶去见周穆王，让它为大家献歌献舞，如图1-2所示。歌舞表演得非常成功，观众陶醉其中。只是歌舞结束的时候出了差错，人偶向穆王的姬妾们“暗送秋波”，惹得穆王大怒，要处死偃师和这个人偶。偃师慌忙解释，并亲自动手把人偶拆卸成一堆零件。穆王这才相信这个人偶确实不是真人，赦免了偃师。鲁班和墨子听说这个故事后大为震撼，从此再也不敢自认为手艺高超了。这个故事显然只是传说，2000多年前是不可能造出如此惟妙惟肖的人偶的。尽管如此，这个故事依然充分体现了古人对智能机器的向往。

图1-2　偃师制造的人偶表演示意

② 早期的自动化机器

对于智能机器的渴望不仅存在于传说中，在很早以前，科学家就开始尝试制造一些自动化机器，帮助人们做事。古希腊数学家、物理学家、发明家阿基米德就制造了许多机械设备，比如阿基米德螺旋提水器、阿基米德之爪、投石机等。这些设备具备一定的自动化能力。阿基米德之爪类似一个带有抓钩的起重机，可以抓起进攻的船只，然后将它摔得粉碎，威力巨大。

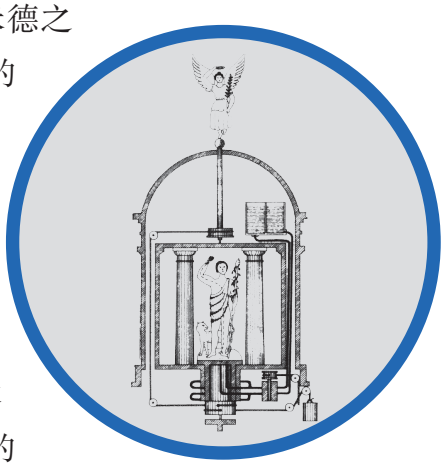


图1-3　希罗的剧院

1世纪，亚历山大里亚的著名数学家兼工程师希罗在他的《自动装置的制作》一书中，描述了一个自动化的木偶剧院（图1-3）：通过轮轴、杠杆、滑轮和车轮等设备之间的互相作用，就可以上演一出完整的木偶剧。自动化木偶剧院的想法受到很多人的推崇，例如，著名的发明家莱昂纳多·达·芬奇不仅在绘画和科学领域有着巨大的贡献，他还设计了许多机械装置和自动化木偶剧院的图纸。虽然达·芬奇的一些设计没有被实际制造出来，但他的构想激发了后来的发明家和工匠。

加扎利（1136—1206）是伊斯兰黄金时代（中世纪）的一位杰出的博学家，集发明家、机械工程师、工匠、艺术家、数学家和天文学家于一身。加扎利制作了很多非常有趣的机器，被誉为现代工程之父。比如他曾制作了一台自动提水的机器，可以把水从水井中提出来。他还制作了一个以水为动力的玩偶乐团（图1-4），通过水流冲击叶片，推动轮轴转动，进而带动连杆上下运动，使玩偶拨动琴弦或敲击鼓面。

图1-4　加扎利发明的自动玩偶乐团

1739年，雅克·德·沃康松制

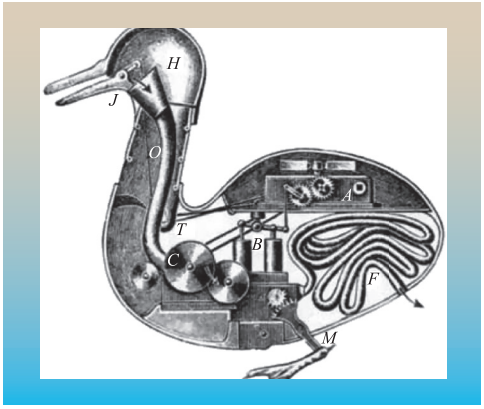


图1-5　雅克·德·沃康松的消化鸭

作了一只“消化鸭”(图1-5):两个鸭翅膀各动用了超过400个活动部件,鸭子体内还有橡胶材质的管道系统。这只鸭子能吃、能喝、能排便。然而,这只鸭子并不具备消化能力,排出的粪便实际上是从暗格喷出的预先制作的丸状物。这些复杂的自动化机械展示了当时工匠们的高超技艺,但其智能程度还是非常有限的。

③ 现代电影中的机器人

进入20世纪以后,随着科学技术的发展,智能机器和机器人的概念在科幻小说和电影中广泛出现。1927年的德国电影《大都会》中的机器人可能是屏幕上出现最早的机器人形象。故事发生在2026年的未来城市“大都会”。大都会的管理者仿照一位名叫玛利亚的女子制造了一个机器人(图1-6)。机器人玛利亚不仅具有人的外貌,还有思想和决策能力。2001年,这部电影被联合国教科文组织列为世界记忆项目(Memory of the World Programme)。

图1-6　电影《大都会》中的机器人形象

20世纪60年代,电影《2001:太空漫游》中出现的超级计算机HAL 9000(图1-7)是智能机器的经典形象。HAL 9000能够与人类对话、理解命令,并且展示出一定的情感和自主决策能力。

进入21世纪以后,智能机器的梦想越来越接近现实,人们也赋予机器人更多情感上的寄托。2008年的电影《机器人总动员》就塑造了瓦力(图1-8)和伊娃这两个情感鲜明的机器人。瓦力原本是人类出走外太空避难时留在地球上捡垃圾的机器人。在漫长的垃圾处理生涯中,因为接触人类的各种物品,渐渐产生了感情,并在偶然的机会捡到了一株植物。有一天,从太空避难所来了一位监工

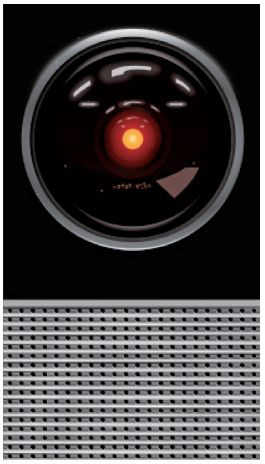


图1-7　电影《2001：太空漫游》中的HAL 9000智能机器

图1-8　《机器人总动员》中的瓦力

机器人，这个机器人叫伊娃。伊娃带走了这株植物，触发了人类回归地球的信号。然而，避难所的自控系统坚决反对人类返回地球，认为那里依然不适合人类生存。瓦力、伊娃和胖船长合作，最终关闭了自控系统，带领人类回到了故乡地球。

4 从梦想到现实

自古以来，人类对智能机器的渴望从未停止。古代的传说和中世纪的自动化机器只是人类梦想的一部分，而电影中的智能机器人则让人们对未来充满了期待。人类希望通过制造智能机器来减轻劳动负担，甚至替代人类完成一些高难度和危险的任务。这种对智能机器的渴望，激励着科学家和工程师们不断进行研究和创新。

然而，梦想终究要面对现实。传说和电影不足为据，那些精巧设计的机械装置虽然在某些方面表现出色，甚至极大减轻了人们的劳动强度，但和人们设想的智能机器相差甚远。人们逐渐意识到，应该寻找一条不一样的路，从根本上解决智能机器的问题，即让机器具备类似人类的思维能力。这就是本书的主题，即人工智能。

思考与讨论

当“消化鸭”刚被发明出来时，大家都觉得这个机器鸭子非常智能。后来，一些人了解了它的工作原理，说它的发明者沃康松是个骗子，鸭子肚子里只不过是一堆机器零件。你怎么看？沃康松是个骗子吗？

课程实践

查找资料，看看历史上或电影里还有哪些你觉得有趣的智能机器，写一篇200字左右的短文，介绍给同学们。

1.2 什么是人工智能

学习目标

（1）了解人类智能的主要类型。

（2）理解人工智能的定义及其通过计算模拟人类智能的核心思想。

我们经常听到人们在讨论人工智能，但究竟什么是人工智能呢？会自动控制温度的冰箱算吗？会定时关闭的电饭煲算吗？本节将澄清人工智能的基本概念，并讨论它与相关学科的关系，探索其强大的根源。

① 智能机器不等于人工智能

人们很早就设计了许多看上去十分“智能”的机器，如阿基米德的提水器（图1-9）、富尔顿的蒸汽船、达盖尔的相机、福特的汽车、会打字的打印机等。这

些机器刚出来的时候都让人非常震惊，因为它们的自动化程度颇高。在我们生活中，这种“智能”机器随处可见，比如会摇头的风扇、会控温的冰箱、会自动洗衣服的洗衣机，商家经常在它们的名字前面加上“智能”二字。这是由于人们倾向于将新颖的、自动化的能力称为“智能”。

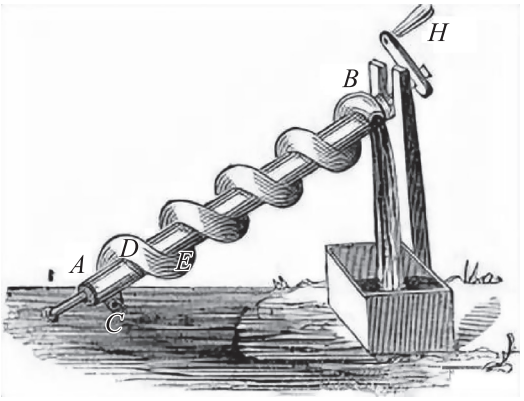


图1-9　阿基米德的提水器（公元前234）

但这并非是真正的智能。所谓的智能行为，包括思考、学习、创造、想象等。人工智能不同于自动化机器，它从最初就是要模拟人的智能，制造“像人一样智能”的机器。这是一个大胆的想法，自动化的机器无论多么强大，都只是工具，按部就班地执行人类设计的工作流程，而模拟人类智能的机器则可以有无限的可能，甚至超越它的创造者。

② 人类智能

要想模拟人类的智能，首先要知道人类的智能有哪些。广义上，任何需要大脑参与的活动都可以认为是智能的（图1-10），如感知、动作、推理、学习、规划、决策、想象、创造、情感等。在这些智能活动里，有些比较基础，如感知、动作；有些比较高级，如推理、想象、创造。不论是哪种智能，都不是可以轻易完成的，都需要大脑的参与，伴随着高级的思维活动。

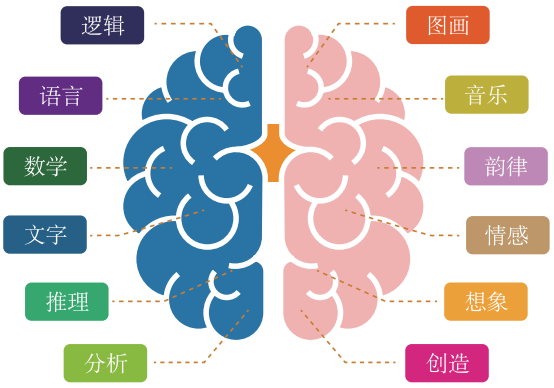


图1-10　大脑左右半球负责不同的智能

值得注意的是，人类的某些基础智能也存在于动物身上，如感知，很多动物的感知能力甚至比人类灵敏；再比如动作，很多动物的速度和灵敏度都超过了人类。但即使是这些基础的智能，人类也与其他动物有明显区别，从听觉感知为例，人类可以

感知到同伴发音中的细微差异，从而通过声音进行交流。人们甚至可以通过声音联想到概念，比如听到“香喷喷的鸡腿”时会流口水，动物则极少有这种能力。再比如动作，人类可以握笔写出漂亮的字，可以完成复杂的雕塑作品，这些都是动物无法比拟的。就算是最简单的行走，人类的直立行走也比动物四足着地行走要复杂得多（图1-11）。总之，人类的这些智能是高级思维能力的体现。人工智能正是要模拟人类的这种高级思维能力。

从这个角度看，汽车四个轮子在地上跑算不上智能，但四足机器狗跑动起来就显得更智能；双脚着地在地板上移动的机器人算不上智能，但在高低不平的路面上自由行走，跌倒了还能爬起来的机器人就非常智能；能听到声音算不上智能，能听懂人类的语言才叫智能。

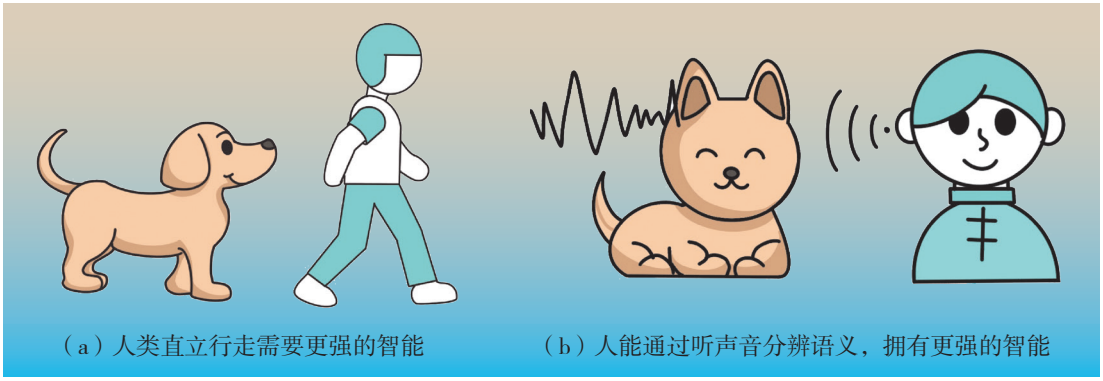


图1-11　人和动物所表现出的智能不同

③ 人工智能的定义

有了模拟人类智能的目标，又该如何实现呢？人工智能的学者们选择的是一条独特的路：通过计算来实现智能。

这一思路最早源于古希腊哲学家亚里士多德对人类思维过程的深刻思考。他认为，人类之所以能获得理性，是因为其思维过程符合特定的规律，符合这些规律的思维过程都被认为是无可置疑的。正是基于这种无可置疑的思维过程，人类才能建立起坚实的知识体系。人类的思维规律被称为逻辑，亚里士多德所

建立的关于思维规律的学问被称为形式逻辑。形式逻辑的确立表明人的思维过程是有规律的。因此，如果让机器依据这些规律进行思考，就可以获得类似人类的智能。这是人工智能发展的源头。后来英国数学家乔治·布尔将形式逻辑表示为符号的演算，这为机器模拟人类思维提供了具体的方案，即逻辑演算。1946年，第一台通用电子计算机 ENIAC诞生，为模拟人类思维提供了实际的计算工具。1956年，约翰·麦卡锡、克劳德·艾尔伍德·香农、马文·闵斯基等在达特茅斯召开研讨会。在这次会议上，由麦卡锡（图1-12）提出的“人工智能”一词成为新科学的名字，人工智能由此登上历史舞台。

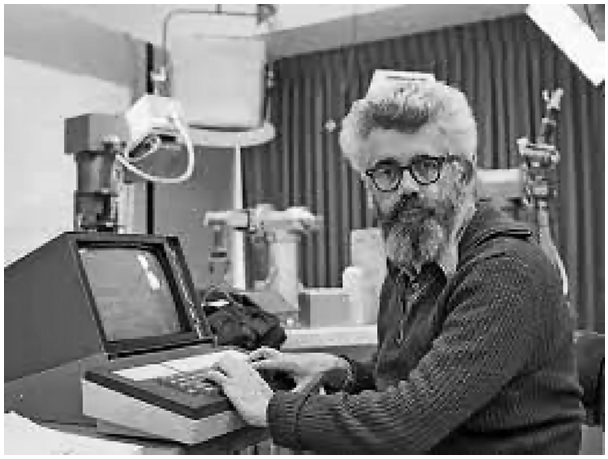


图1-12　约翰·麦卡锡（1927—2011）

从上述人工智能的诞生过程可以清楚地看到，人工智能起源于对人类思维的模拟，采用的方法是数学计算，工具是计算机。后来，人工智能的研究从狭义的“思维”（thinking）扩展到广义的智能（intelligence），包括感知、学习、情感等。人工智能的研究者相信，只要这些智能过程可以表示为计算过程，就可以被机器所模拟，最终实现类似人类的智能。

目前，实现人工智能主要有两种方案，一种是让机器模拟人类的智能行为，一种是让机器模拟人类大脑的工作机理。目前，模拟智能行为是当前研究界的主流。这里的“智能行为”指智能的外在表现，包括感知、动作、推理、学习、规划、决策、想象、创造、情感等。因此，我们定义“人工智能”是用计算机模拟人类智能行为的科学（图1-13）。

图1-13　人工智能的定义示意

这个定义可以帮助我们澄清很多模糊问题。比如，一台会自动跳闸的电饭锅是人工智能吗？显然不算，因为它依据的是“当温度过高时，磁铁的磁性消失”这一物理现象，而自动跳闸不是通过计算实现的。然而，如果电饭锅里具备了更强大的功能，比如可以自动学习如何做饭更好吃，那就包含了人工智能的成分。再比如一台计算器算不算人工智能呢？从功能上看是的，因为数学计算确实是人类的重要智能，而且这一智能在计算器里也确实是用计算的方式实现的。然而，它的模拟过程和人类大脑里的实际处理过程相差甚远，很难作为典型的人工智能实例。这也说明一些智能活动对人类来说比较困难，对机器来说反而很简单，类似的还有记忆能力、运动能力等，都是机器比较擅长的。

总体上看，人工智能更加关注需要人类高级思维能力才可以完成的任务。在人工智能发展初期，研究者更关注比较基础的智能，比如我们前面提到的识别人脸、识别声音、让机器开口说话等。随着技术的进步，人工智能更关注高级智能，如推理、创造、决策等，例如，以ChatGPT为代表的大模型已具备很强的推理和创造能力。随着人工智能的发展，用计算模拟人类高级思维能力的学科特点愈发明显。

④ 人工智能的普适性

人工智能是一门非常特殊的学科，这种特殊性源于对人类智能的模拟。所有学科都建立在人类智能的基础上，数学和物理也不例外。如果人工智能可以模拟人类智能，这意味着它可以应用在任何领域，甚至可能开创新的学科。

近年来，人工智能在生物、物理、化学、天文等各个领域都取得了令人瞩目的成就，也反映了这一趋势。一方面，这是因为这些学科已经发展了很久，传统方法已经很难继续创新，而人工智能带来了新的思维方式和研究手段，特别是基于大数据的学习方法，成为各个学科进一步发展的巨大推动力；另一方面，这也说明人工智能正在成为超越学科边界的基础工具（图1-14），不论将来从事哪方面的工作都必须用到人工智能。这也是我们要认真学好人工智能的主要原因。

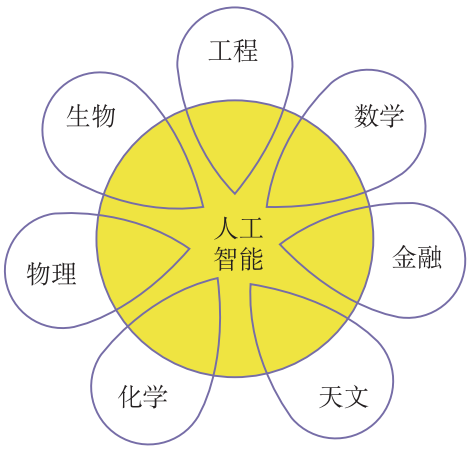


图1-14　人工智能与基础学科的交叉融合

5 总结

什么是人工智能？这是所有人工智能的初学者问的第一个问题。这个问题不易回答，因为历史上为了实现智能机器的梦想，人们提出了很多方法和思路，这些繁杂的方法和思路容易淹没学科的主线。尽管如此，如果把视野放到历史纵深，就可以看到人工智能的独特之处：它起源于对人类思维的模拟，采用的手段是计算。这是人工智能学者给自己选定的一条充满荆棘的道路，也正是这条特殊的道路，造就了今天人工智能这门学科的辉煌。

目前，实现人工智能的主流方法是用计算机模拟人类的智能行为，即模拟人类智能的外在表现而不是复现人类的智能过程本身。这是因为人类智能的生理过程非常复杂，要复现这一过程目前还很困难，然而模拟人类智能的外在表现（包括我们的思维规律）还是可能的，也是可以检验的。当然，人工智能领域的科学家也不会拒绝生理学和神经学对智能本身的研究成果，而是会和这些领域的科学家们通力合作，将最新研究成果应用到人工智能的理论和实践中。例如，人工神经网络就是受人类神经系统，特别是大脑工作机理的启发而设计的计算模型，目前已经成为人工智能中最重要的方法。

人工智能是模拟人类智能行为，特别是高级思维能力的科学，这意味着它必然渗透到其他学科中，形成深入的交叉融合。因为所有学科建立和发展的基础都是人的智力活动，当人工智能足够强大时，理论上它可以代替人类推动各门学科的进步。

思考与讨论

有人说：人工智能的背后是数学，数学是人工智能的灵魂。结合本节中关于人工智能的历史起源，说说你对这一观点是否同意。

课程实践

组成研究小组，选择一门基础学科（如物理、化学、生物、数学），查找资料，看看当前人工智能方法在该学科中是否有应用，应用在哪些方面。做一张海报来报告你们的调研结果。

1.3 机器的眼睛

学习目标

（1）认识机器视觉的基本概念及其重要性。
（2）理解人脸识别、车牌识别和物体识别的基本原理与应用场景。
（3）探讨机器视觉的积极影响与潜在风险。

眼睛是人重要的感知器官，因为有了眼睛，我们才可以看到丰富多彩的世界，才能轻松地在其中行动和探索。不难想象，如果没有眼睛，我们的生活将变

得多么艰难。同样，给机器装上眼睛，让机器拥有强大的视觉能力，是制造智能机器的首要目标。

机器拥有眼睛以后可以做很多事，比如分辨人脸、识别车牌、辨认红绿灯、发现火灾、预测抛物轨迹等。理论上，任何人眼所能做的事，机器的眼睛也一样可以做到，甚至比人做得更好，比如从太空望远镜拍摄的图片中发现新星，从病理切片的显微图片中发现病灶区域等。这方面的研究统称为“机器视觉”（computer vision）。这是个庞大的研究领域，本节通过几个日常生活中常见的应用来展示机器视觉的强大功能。

① 人脸识别

识别人脸是人类视觉的重要功能。人类是如何识别人脸的呢？首先眼睛接收到包含人脸的图像，并将其传入大脑的枕叶区，检测出人脸区域，再传递到梭状回这一特殊脑区，完成面孔的匹配与辨识。

机器视觉也采用类似的方法，首先检测面部区域，再判断是否认识这张脸。人脸识别的研究开始于20世纪60年代，当时的思路是通过标记人脸的关键点并提取典型征来进行识别，如标记两眼之间的距离，嘴唇的大小等（图1-15）。这种方法虽然有一定效果，但性能较差，因为关键点和特征的提取不稳定，特别是在光照、角度等变化时，性能会显著下降。后来，研究者提出了多种方法，但始终无法满足真实场景下对识别精度的要求。直到2014年，深度神经网络的出现大幅提升了人脸识别的精度，尤其在复杂场景下的表现有了显著改善。此后，人脸识别才真正走进我们的日常生活。目前，人脸识别的精度极高，例如，在LFW（labelled face in the wild）人脸数据集（图1-16）上，最新方法的准确率已经可以超过99.8%。

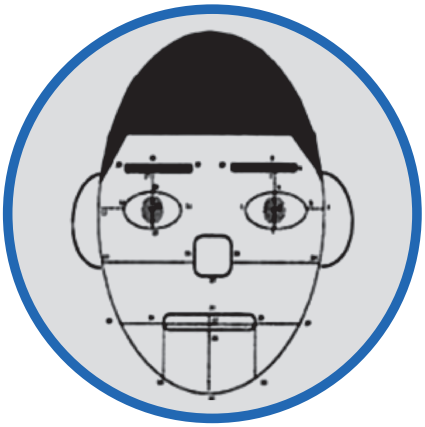


图1-15　用于人脸识别的几何特征

图1-16　LFW人脸数据集

如今，人脸识别技术在很多领域被广泛应用。例如，人脸支付可以通过识别人脸完成支付过程，不需要携带现金、信用卡，或手机扫码，极大提升了便捷性。人脸识别还被用于机场安检、高铁刷脸检票、超市购物（图1-17）等场景，在保证安全性的同时极大提高了通行效率。警方还利用人脸识别抓捕通缉犯。将通缉犯的照片输入监控系统后，一旦嫌犯出现，就可能被布置在各处的摄像头检测出来并触发报警。据报道，这一系统已经协助警方抓获了多名潜逃多年的通缉犯，战果辉煌。更值得一提的是人脸识别技术还被用于宠物识别。例如，把宠物狗的照片上传到数据库，这样便于在它走丢时及时寻回。

图1-17　超市的人脸支付

② 车牌识别

车牌识别是计算机视觉的另一个典型任务。车牌识别技术最早由英国警察研究机构于1976年开发的，用于打击车辆盗抢活动。最早的实验系统安装于英国的A1公路，并于1981年首次定位并找到了一台被偷的车辆。然而，由于成本较高且精度不足，这一技术早期并未广泛使用。20世纪90 年代，伦敦的“钢环”计划推动了摄像头网络的建设和自动车牌识别数据中心的建立。2010年以后，随着数据量的积累和人工智能新方法的引入，车牌识别精度大幅提升。目前，在正常环境下，车牌识别的精度可达99%以上，几乎没有误差，使得这一技术可以大范围应用。如今，全球许多国家已在公路上部署自动车牌识别系统，代替警察监控交通违法行为。这些系统被形象地称为“电子交警”。

车牌识别主要包括两个步骤：车牌定位和内容识别（图1-18）。车牌定位是确定车牌区域，而内容识别是识别车牌里的数字和字母。在实验室条件下，不论定位还是识别都问题不大，但在实际应用中就没有那么简单。这是由于现实环境中存在多种复杂情况出现，导致性能下降。例如，车速过快时，定位与识别精度会下降；当车流量过大时，识别所有车牌也是一大挑战；雨雪天气下，光照条件差导致图像质量下降，识别难度进一步增加。2010年后，基于深度神经网络的新方法极大提高了在这些复杂场景下的识别性能。电子交警不仅可以识别车牌，还可以识别吸烟、接打电话、未系安全带等违章行为。可以想象，如果没有这些系统，现代城市中车流如织，警察可能难以应对所有违章行为，社会稳定也将面临挑战。

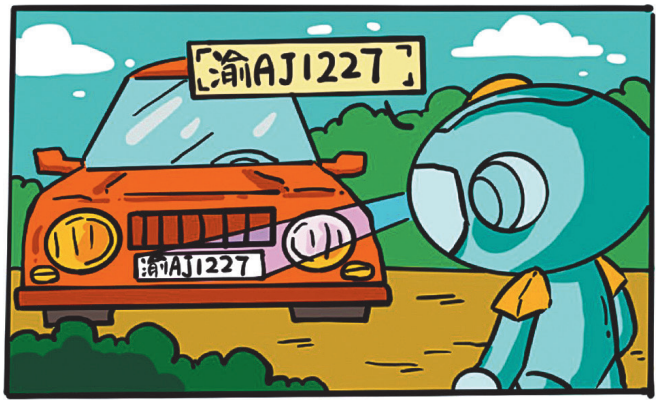


图1-18　从图片中定位车牌位置并识别车牌内容

③ 物体识别

自动驾驶汽车需要识别前方是一辆车还是行人，需要从视野中发现红绿灯，还需要识别地上的斑马线。这背后的技术称为物体识别，即将图片中的物体框出来并标注类别。如图1-19所示，图片中包含汽车、卡车、自行车、人、红绿灯等物体，物体识别系统需要为它（他）们标注不同的类别标签。

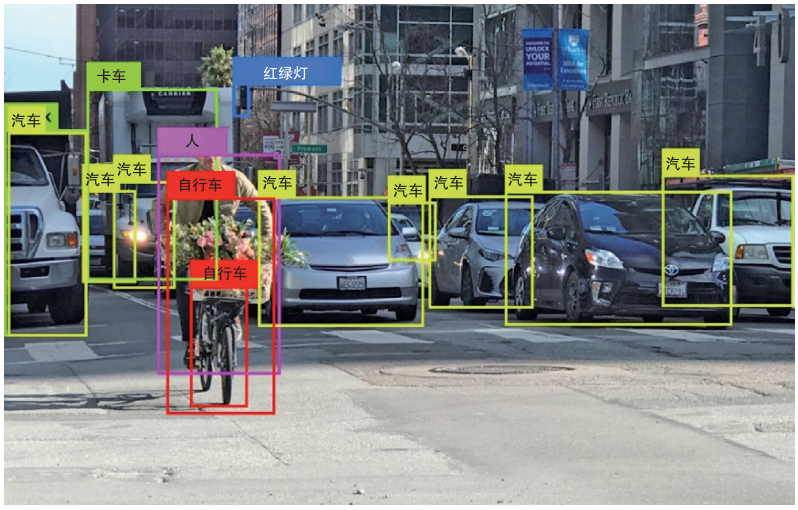


图1-19　物体识别

物体识别是人类视觉的基本功能，可以认为是对场景的初步理解。早在20世纪60年代研究者就希望计算机能像人类一样理解图像中的内容。1982年，科学家大卫·马尔提出了一种分步骤理解图像的方法：先识别整体形状，再分析细节。这些早期研究都无法深入理解图像里的内容，因此性能都不理想，特别是在场景中出现大量复杂物体时更难以识别。2009年，华人科学家李飞飞发布了大规模图像数据库ImageNet（图1-20），包含320万张实际场景图片，按12个亚类、5247个子类进行标记（目前已达到1400万张图片，21841个子类）。2010年，基于ImageNet的物体识别竞赛启动，当时最佳的系统的ToP-5错误率（真实类别不在系统预测的前5个类别中视为一个错误）是28%，可见这一任务的困难。2012年，基于深度神经网络的新方法，将识别错误率大幅降到16%。此后，物体识别的精度越来越高，到2017年，错误率已降至2.25%，超越了人眼的识别精度。

图1-20　李飞飞和她的ImageNet数据库

物体识别具有广泛的应用场景。除了前面说的自动驾驶，物体识别还可以用于机器人环境感知，以及在无人商店中识别用户购买的物品。比如，亚马逊推出的无人商店amazon go（图1-21），无须店员值守，摄像头会自动识别顾客选购的商品并计算总价。客户完成购物以后不需要结账，可以直接离开，系统会自动从其账户中扣款走。

图1-21　无人商店amazon go

④ 结论

让机器拥有视觉能力，使其能够“看到”世界，是人工智能学者们长久追求的目标。随着技术的进步，这一目标已成为现实，机器不仅具备了视觉能力，而且在很多重要场景里已经超越了人眼的精度。比如，在人脸识别和车牌识别中，机器视觉的精度已经超过99%；在物体识别中，精度超过97%。不仅如此，机器视觉在众多领域也都表现出色。例如，探查金属结构的细微变化、从病理图片中识别病灶（图1-22）、从显微镜图片中发现炭疽病菌、检测伪造图像中的微小差异等。这些任务都超过了人类视觉的能力范围。

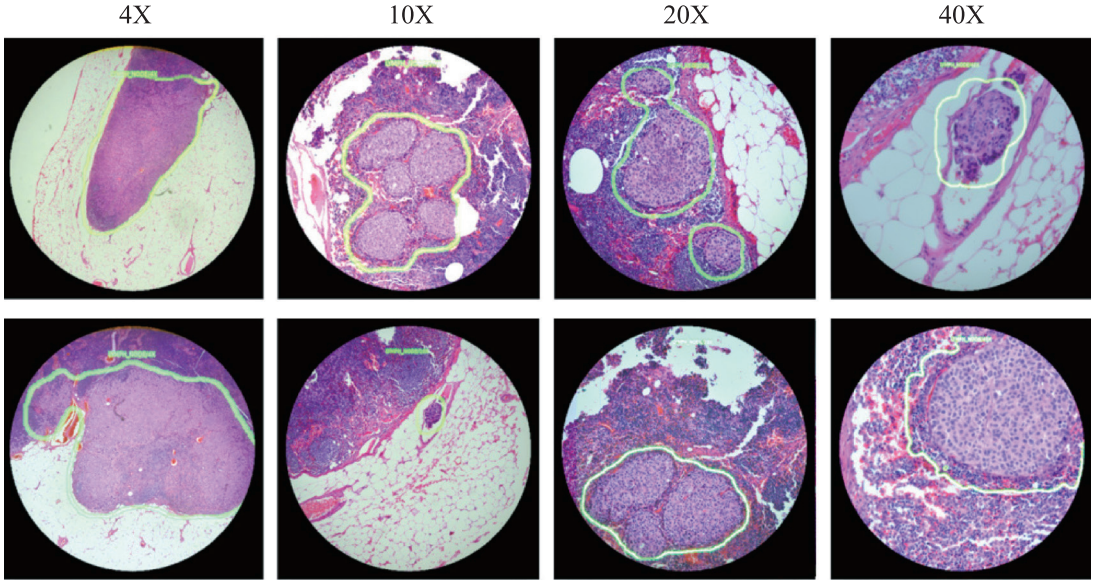


图1-22　谷歌AI增强显微镜自动标出病灶区域

此外，机器正通过其视觉系统观察和理解世界。对于机器而言，视觉系统仅是感知世界的窗口，更重要的是对视觉信息的分析与理解。比如，OpenAI的GPT以及其他大模型工具已经可以基于图片内容进行“看图说话”了：上传一张图片，问GPT一个关于图片的问题，它通常能给出准确的回答。此外，一些科学家正尝试让机器通过观察分析物体属性（冷热、软硬等），甚至自动总结出未知的物理规律。

值得注意的是，机器视觉可能与人类视觉存在差异。换言之，机器感知世界的方式可能与人类不一样。这并不奇怪，因为人类视觉基于生理结构，而机器视觉则有其独特的技术基础。这种差异性意味着机器可能会超越人类视觉的极限，获得更强的视觉能力，例如，从浩瀚的星空中发现一颗变化的新星。然而，这种差异也可能带来潜在风险：人类和机器合作的前提是对场景有一致的判断，如果缺少这种一致性，机器所做出的行动和决策可能与人类的预期不符。比如我们看到的是一个苹果，机器可能将其识别为一团火，进而错误地使用灭火器喷射苹果。目前，机器与人类在视觉感知上的差异并不大，但潜在风险依然存在。如何确保机器和人类在视觉感知上保持一致，是机器视觉研究的重要课题。

1.4 机器的耳朵

学习目标

（1）了解机器听觉的基本概念及其研究方向。

（2）理解语音识别、声纹识别和声音事件检测的基本原理与应用场景。

除了眼睛，耳朵也是最重要的交流器官之一。我们都有这样的体验：闭上眼睛，仅凭听觉也能感知周围环境并了解正在发生的事情。因此，给机器装上耳朵，同样可以使其更加智能。研究如何让机器感知声音的学科称为机器听觉（computational auditory），既包括识别人类的声音，也包括感知自然界的声音。本节将探讨机器如何倾听世界。

① 语音识别

理解他人语言是耳朵最重要的功能之一。人类能轻松通过声音进行交流并

理解对方的语言，这是经过长期进化获得的能力。然而，让机器理解人类语言并不容易，因为人们的发音系统极为精巧，非常细微的发音差异即可区分不同含义，如汉语的“猪”和“煮”，发音很相近，但意义完全不同。此外，相同发音在不同上下文中可能含义也不同，如英语里的to、too 和 two。机器要分辨这些细微的发音差异很困难。另外，人的语言中包含着大量信息，如果缺乏相关知识，很难理解对方在说什么。同样，机器需要掌握大量背景知识才能有效识别出语音内容。

从声音信号中提取发音内容的任务称为语音识别。我们知道声音是声源的振动在空气中进行传播所产生的波动。对于人类语音，声源就是声带的振动。从物理角度看，机器接收的声音只是一长串的振动信号，要想从这些振动信号中将发音内容提取出来非常困难。早期的语音识别仅能区分很简单发音单元，比如贝尔实验室的AUDREY系统，可以识别十个数字。随着技术进步，语音识别的性能显著提升，研究者设计出多种声音信号表示方法（称为特征），用更合理的模型来描述语音生成过程，并引入语言知识别更准确。近十年来，随着海量数据的积累和以深度神经网络为代表的强大识别模型的出现，机器的听觉能力显著提高。2016年，微软的研究人员宣布其语音识别系统在电话语音识别任务上错误率仅有5.9%，超过了人类听音员的水平。

近年来，语音识别在多种复杂任务中取得了令人振奋的成果。目前最强大的语音识别系统是由OpenAI公司发布的Whisper，该系统使用了68万小时的语音数据，训练模型中2/3是英语，1/3是非英语。它不仅可以识别出发音的内容，还可以将其他语言的发音直接转写成英语。实验表明，Whisper在复杂数据集上的表现达到或超过人类专业标注人员的水平。2024年5月，OpenAI推出的GPT-4o支持50多种语言的对话交流，其发音的自然程度令人震惊。在国内，百度、阿里、腾讯、科大讯飞等都推出了自己的中文语音识别产品（图1-23），其在准确度上已

图1-23　智能音箱可以与人通过语音自然交互

经可以满足日常交流的需求。

语音识别技术的广泛应用显著改变了人们的生活方式，使人机交互变得更加自然和便捷。例如，现在很多人已经习惯用手机里的语音转文字功能口述信息，而不是用手写或拼音输入。家里的智能音箱、扫地机器人、抽油烟机等也可以支持语音命令，非常便捷。在开车时，手动操作导航设备会增加交通安全隐患，而语音控制不仅降低了风险，还提升了驾驶体验。此外，语音识别也为残障人士提供了便利，使他们能够更轻松地与他人互动。例如，带有语音识别能力的轮椅可以让行动困难的病人扩大活动范围，视力障碍者也可以通过语音控制家里的智能家电，提高生活质量。

② 声纹识别

人类的耳朵除了用于听懂发音内容之外，还可以识别出谁在说话。每个人的说话方式都是独特的，世界上没有哪两个人的声音是完全一样的。这与指纹类似，每个人的指纹都是独一无二的。借鉴指纹的概念，声音中独特的说话人特性被称为“声纹”。实验表明，人的耳朵有很强的听声辨人能力。我们都有这样的体验，接电话时，如果对方是熟人，只需要“喂”一声即可识别对方的身份（图1-24）。有趣的是，对于陌生人，人耳的判断能力会显著下降。比如对比两段陌生人的声音，我们往往很难判断它们是否来自同一个人。

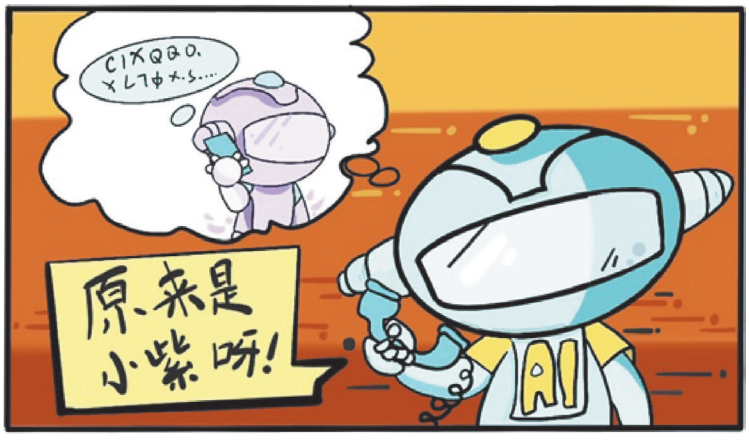


图1-24　声纹识别可以通过声音判断人的身份

声纹识别，学术上称为“说话人识别”，旨在让机器通过声音自动识别发音人身份。阿拉伯故事《一千零一夜》中的“芝麻开门”以及很多科幻电影中都有类似的场景，都体现了用声音验证身份的概念。事实上，对这一技术的研究早在20世纪70年代就开始了，但是性能一直无法达到实用的程度。这是因为说话人特征具有很强的变动性：环境、说话方式、语气甚至姿势的变化都会引起发音方式的变化。这意味着声纹虽然具有个体唯一性，但却不是一个固定的特征，而是随着情景变化的，这与出生就保持基本不变的指纹截然不同。更糟糕的是，同一个人的声纹变化甚至会大于不同人之间的差异，这表明声纹信息天然具有模糊性。

近年来，深度神经网络被引入声纹识别领域，取得了显著成功。这一方法的基本思路是通过层次性信息处理，保留与说话人相关的特征，逐步去除与说话人无关的特征，从而在保持说话人区分性的同时，减小同一个人内部的变动性。目前，最先进的声纹识别系统在VoxCeleb的测试集上可以实现低于1%的错误率，但在更复杂的CNCeleb测试集上错误率仍高于5%。这样的性能已经可以在安全性要求不高的场景中使用。

例如，智能家居可以通过识别家庭成员的身份，提供个性化的服务。比如，智能音箱检测到“播放歌曲”的指令是由一个孩子发出的，它将倾向于播放儿歌。此外，通过声音比对，可以筛查嫌疑犯，缩小追踪范围，加速案件侦破。电影《燃眉追击》中有这样一个场景：一位听音专家通过一段录音判断说话人的特征为“古巴人，35～45岁，在美国东部受的教育……”，随后这段录音被送到一台超级计算机中和一个嫌疑人的视频做比对，发现嫌疑人是罪犯的可信度为90.1%。这一略带夸张的故事情节反映了人们对声音技术在社会安全领域的期待。

然而，在某些关键场景中应慎用声纹识别。例如，仅凭声音分析给嫌疑人定罪就可能面临极大风险。事实上人们很早就将声音分析应用在司法审判中，比如提供一些分析工具帮助专家比对声音，以判断案发现场的人员身份。后来，声纹识别系统也被应用于司法实践，机器可以自动识别声音样本的发音人并给出概率。不论采用哪种方式，都可能面临巨大风险。

最典型的一个案例是波普案。1986年大卫·肖恩·波普在得克萨斯州因强奸罪被定罪，定罪的重要依据之一是休斯敦警察局向陪审团展示了受害人电话录音与波普的声音的比对，声称这两者具有相似的模式。波普在服刑15年后，于

2001年通过DNA检测证实了他是清白的。另一个案例发生在法国，警方接到一名男子的电话，其声称要为一起汽车炸弹袭击负责。后来一个叫Jerome Prieto的人被发现与该电话里的人发音一致，导致他受到10个月的非法羁押。此案过后，法国声学学会发布声明，要求停止声纹技术在法庭上的应用。这些案例表明，声纹可以作为辅助信息帮助法官判案，但目前还不能作为主要证据或唯一证据。

③ 声音事件检测

人耳不仅能感知声音，还能识别大自然的声音，比如鸟鸣、雷声、雨声等。智能机器要想更好地感知世界，要有能力从混杂的声音中提取出各种不同的声音事件，这一任务称为声音事件检测。声音事件检测与机器视觉中的物体识别类似，都是从一个场景中（一幅画或一个声音序列）中检测出包含的对象，但声音检测更复杂一些，因为不同声源的声音会互相叠加，要从混叠声音信号中提取事件更具挑战性。如图1-25所示，人的语音和音乐同时出现，而汽车声和音乐声也互相叠加。这些叠加在时间上和音量上都各不相同，形成了复杂的混合信号。2017年，谷歌发布了一个名为AudioSet的大规模声音事件数据集，包含200多万条音频，涵盖了527个类别。目前，人工智能模型在这一数据集上已经表现出了优异的性能。

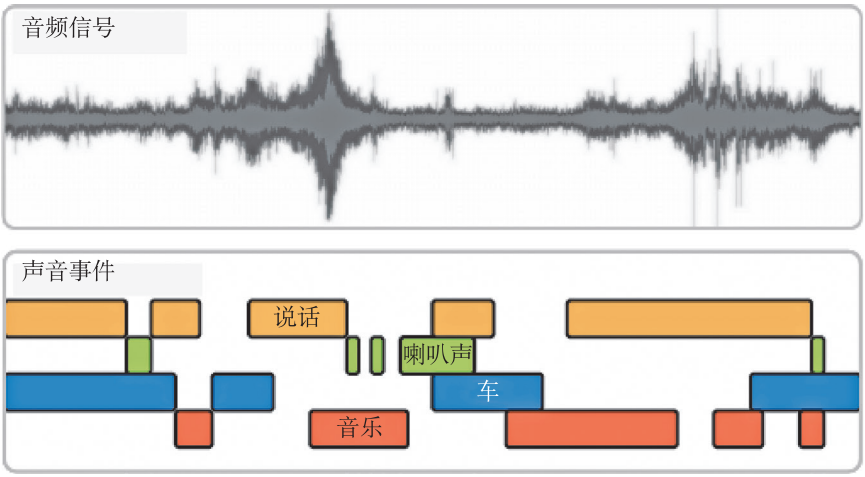


图1-25　从语音信号中检测出人声、车辆、喇叭声、音乐等声音事件

④ 总结

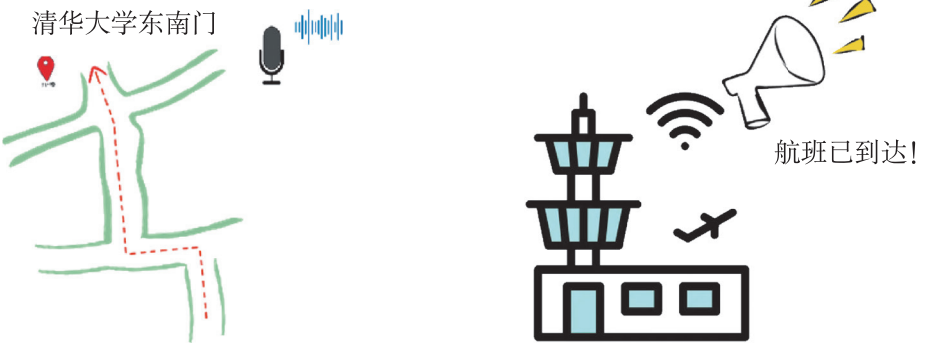
我们介绍了机器听觉中的部分研究内容，这些只是机器听觉的基本功能。除此之外，研究者还在探索如何从声音中判断人的情绪，识别声音的来源，将混合声音进行分离，对声音场景进行分类等任务。这些功能也是人类听觉的常见应用。尽管机器在许多任务上已超越人类听觉（比如声纹识别），但在一些任务上还比不过人类，比如识别情绪，人类可以轻松判断语音中的情绪信息，哪怕对说话内容一无所知，机器目前还做不到这一点。另外，在多人说话的“鸡尾酒会”场景中，人耳可以聚焦并识别特定说者，而机器在这方面还有明显差距。然而，随着技术的进步，很多难题被逐一解决，机器听觉全面超越人类或许只是时间问题。

1.5 机器的嘴巴

学习目标

（1）了解语音合成技术的原理及其发展历程。
（2）理解现代语音合成技术的主要方法。
（3）探讨语音合成技术的风险及其社会影响。

会说话的机器如今已十分常见了，比如送餐机器人能与人类对话，导航软件会实时播报路况（图1-26），新闻客户端还能朗读新闻，然而，如果将任何一种会说话的机器带回到200年前，都会让人无比震惊。这种让机器发声的技术称为语音合成。本节我们将讨论人工智能是如何实现语音合成，让机器开口说话的。



（a）地图导航　　（b）机场广播

图1-26　语音合成技术应用举例

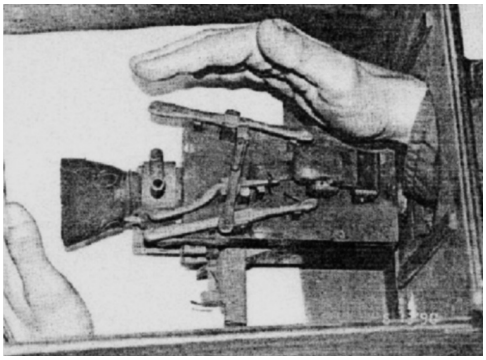
① 早期的机械发声机器

让机器开口说话是人类长久以来的梦想。

真正的会说话的机器出现在1769年。那年，匈牙利发明家沃尔夫冈·冯·肯佩伦依据人类的发声机理，制作了一台机械发声器，这是让机器发声的早期尝试。人类是如何发声的呢？肺部的气流冲击喉部的声带，引起声带振动，振动经过口腔和鼻腔传导后从口唇传递出来，就形成了我们听到的声音。如果可以通过机械的方式把声带的振动以及口腔和鼻腔的传导过程模拟出来，就可以发出类似人的声音了。

肯佩伦的发声机器正是依据这一原理制造而成的。如图1-27所示，皮质的风箱用以模拟人的肺部；木质的空箱内置一个阀门，用以模拟人的喉部和口鼻。挤压

（a）外观



（b）内部构造

图1-27　肯佩伦发声器

风箱时，气流通过阀门进入空箱，从而模拟人类发音。实际操作中，通过调整阀门处的孔洞，即可模仿不同的发音。

后来，人们又依据类似的原理设计了很多机械发音装置。然而，这些机器只能发出一些简单的声音，如“爸爸”“妈妈”，而且声音比较模糊，无法发出连贯清晰的声音。

② 声码器：现代语音学的开端

1939年，贝尔实验室的科学家荷马·达德利发明了声码器，标志着现代语音合成技术和现代语音学的开端。与之前的机械发声装置不同，达德利的声码器是一种电子发音设备，利用计算原理来合成声音。具体来说，它把人的声音拆解为两个环节的前后耦合：一是喉咙的振动，二是口腔和鼻腔组成的传递通道。关键在于，这一耦合可以通过计算实现，这样就可以利用电子电路来模拟声音的产生过程，并通过扬声器将声音播放出来。

1939年，基于声码器原理VoCoder发声器在纽约世界博览会上展出，引起轰动（图1-28）。操作人员通过键盘控制发音内容，用脚踏板调整音高，实现了让人震撼的连续发音效果。

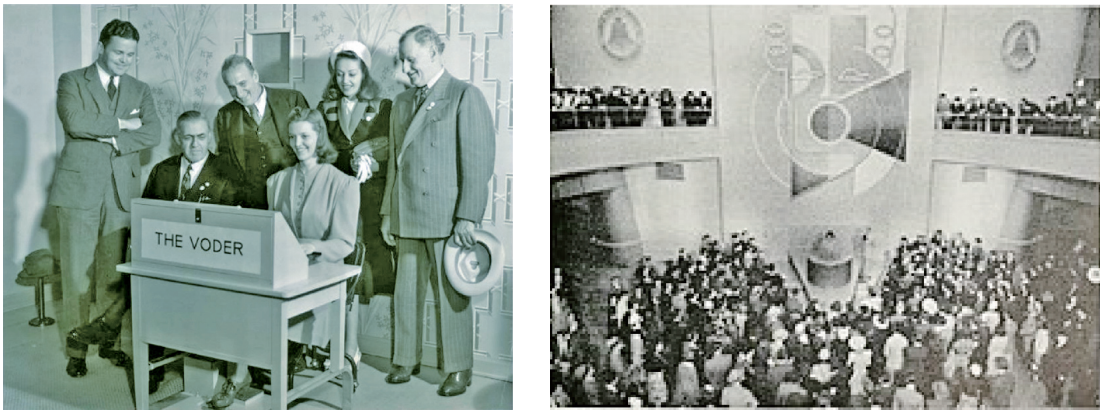


图1-28　贝尔实验室在1939年纽约世界博览会上展出发声器的场景

③ 现代语音合成技术

现代语音合成技术的发展也经历了很多波折。早期的语音合成器完全基于达德利的声码器，它的发音过程是由计算机自动控制而非人为控制。如图1-29所示，它先将句子拆分成发音单元，再为每个发音单元计算出声码器的参数（包括声带振动的频率和口鼻传递声音的特性），最后再交由声码器合成出声音。

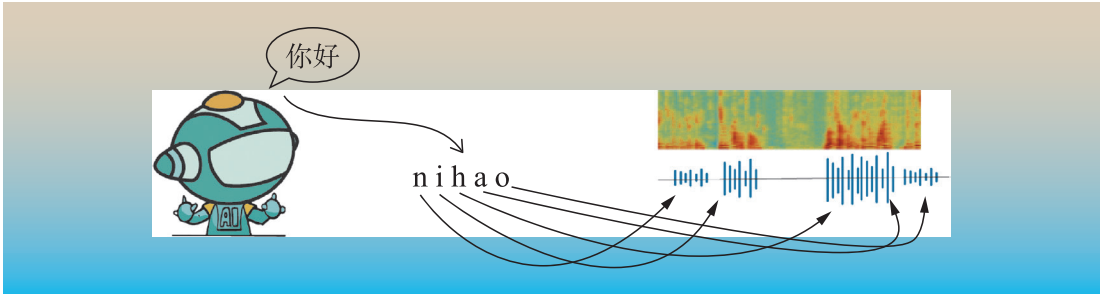


图1-29　现代语音合成技术的步骤

这种合成方法结构紧凑，声音质量也比较清晰，代表产品是DEC公司的DECtalk DTC01（图1-30）。这种发音有明显的机械感，常用于街机游戏。著名科学家霍金的轮椅也曾长期使用这种发音技术。

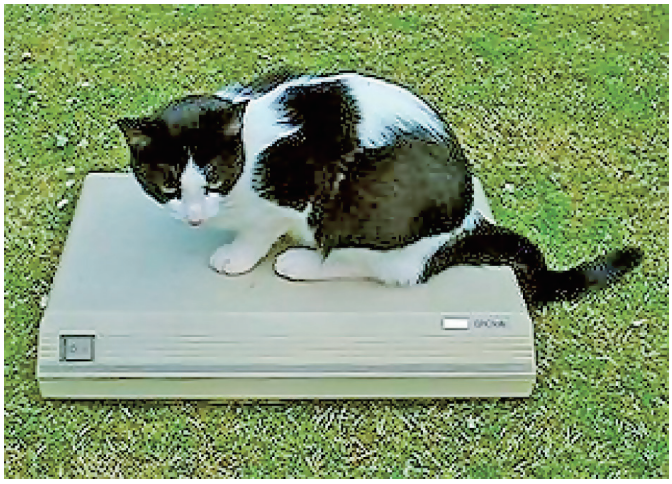


图1-30　加装共振峰的语音合成系统DECtalk DTC01

随着技术的进步，人们渐渐无法接受带有浓厚机械感的声音，希望合成的声音更加自然，最好和人的发音无异。为什么会有机器腔呢？这是因为发音模型过于简单，不能反映发音过程的复杂性。大规模神经网络出现以后，研究者利用神经网络强大的建模能力，基于大数据自动学习人类的发音过程，实现了高度逼真的模拟，发音质量得到极大提高，甚至达到了真假难辨的地步。比如导航软件中的声音自然流畅、风趣幽默，几乎和真人发音没有什么区别。

除了能从文本合成声音，目前的人工智能技术还可以对声音进行转换，比如在保证发音内容不变的前提下将小明的语音，转换成小红的语音。这一转换技术，是通过一个神经网络提取发音内容，去掉小明的发音特征，再经过另一个神经网络，加入小红的发音特征，最终生成小红的语音。此前备受关注的“AI孙燕姿”，就是基于该技术实现的。“AI孙燕姿”用孙燕姿的音色翻唱了众多歌曲，播放量极高，目前至少合成了1000首歌曲。除了“AI孙燕姿”，还有“AI 张雨生”“AI周杰伦”等各路AI歌手，甚至普通人也可以体验当明星的感觉。

④ 语音合成技术的风险

越来越逼真的合成语音也带来了很大的风险。主要体现在两个方面：一是被用于欺诈，二是存在版权风险。

（1）欺诈风险方面。合成语音可能被不法分子利用，冒充他人进行诈骗，导致人们遭受财产损失。同时，语音合成技术还可能用来伪造司法证据，造成冤假错案。更进一步，逼真的合成语音还可能骗过声纹身份验证系统，带来恶意闯入的风险。

（2）版权风险方面。以“AI孙燕姿”为例，一首用AI合成的孙燕姿音色的歌曲上传到网上可能会带来极大流量，但同时，也产生了版权归属的法律难题：这首歌的版权到底应该属于谁？是转换之前的版权所有人，是孙燕姿，还是用AI技术制作歌曲的人？到目前为止这些问题还没有定论。

⑤ 总结

经过多年研究，机器已经可以开口说话了。早期的发声技术用机械模拟人的发音机理，无法生成连续的、可分辨的声音。进入计算机时代后，研究者们设计了人类发音的计算模型，并基于这一模型发明了声码器，使得机器可以发出清晰、连贯的声音，但仍带有明显的机械感，不够自然。得益于语音数据的积累和神经网络的学习能力，目前机器已经可以非常细致地模拟人的发声过程，从而发出自然、逼真的声音。可以说，能听会说的机器已经来到我们身边。

1.6 机器的手和脚

学习目标

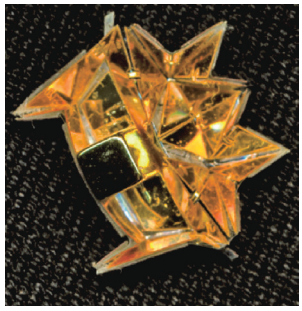
（1）理解实体机器人的定义及其主要类型（移动机器人与动作机器人）。

（2）了解扫地机器人、自动驾驶汽车和机械手的基本原理与实际应用。

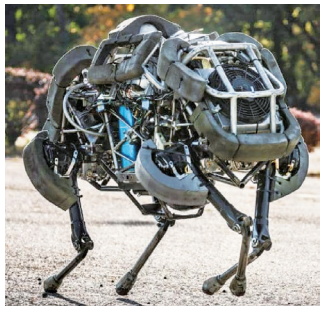
（3）探讨机器人与人工智能的关系。

我们经常在电影里看到各种机器人的形象，比如《超能陆战队》里憨厚可爱的大白，《机器人总动员》里勇敢善良的瓦力。事实上，机器人是一个广义的概念，涉及的范围很广泛，其中既包括手机里的聊天机器人、电话客服机器人等虚拟形态，也包括拥有物理形态的并能通过行动影响周围环境的实体机器人。值得说明的是，机器人未必是人的形状，也可以是昆虫、小车、机械臂等，如图1-31所示。

（a）机器昆虫



（b）可吞入的手术机器人



（c）波士顿动力公司的“大狗”

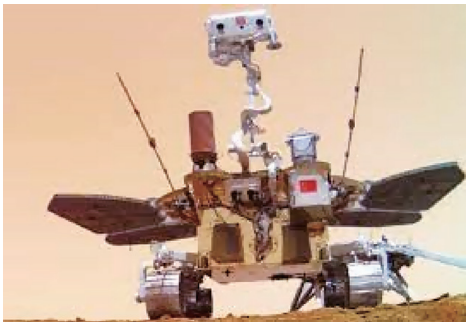
图1-31　各式各样的机器人

按照行为方式，机器人可以分为两种：一类是“移动机器人”，如自动驾驶汽车、无人机等；能够在不同位置间移动。另一类称为“动作机器人”，如工业生产中的机械臂，帮助医生进行手术的手术机器人等，通过肢体动作完成特定任务。移动机器人模拟的是人类的路径规划能力，而动作机器人模拟的是人类的肢体动作能力。在实际应用中，这两者往往结合在一起，比如一台火星车，既要想办法规划路径以接近目标，也需要执行一些抓取、采集任务（图1-32）。再比如波士顿动力公司推出的“大狗”机器人，它要完成一个爬山任务，需要确定好目标，规划好路线，再手脚并用地完成爬山动作。

目前，机器人已经走进了我们的日常生活，不论是家里勤勤恳恳的扫地机器人，还是宾馆里的送物机器人，或是从不违章的自动驾驶出租车，机器人正在成为人类各个领域中的亲密伙伴。本节将探讨人工智能在机器人方面的应用，重点是扫地机器人、自动驾驶汽车和工业机械手的人工智能技术。

（a）自动驾驶汽车

（b）机械手



（c）火星车

图1-32　不同行为方式的机器人

① 扫地机器人

扫地机器人是一种可以自动完成地板清洁工作的机器。1996年，伊莱克斯发布了第一款扫地机器人三叶虫（Trilobite）（图1-33），引发了大量关注，但因为价格昂贵没有得到广泛应用。直到2002年，iRobot的Roomba问世并实现大规模量产，扫地机器人才广为人知。

图1-33　三叶虫扫地机器人

目前，市面上已经出现了各种类型的扫地机器人（图1-34），有的机器人甚至可以爬上墙去擦玻璃。

图1-34　各种形态的扫地机器人

扫地机器人的运动和清扫部件并不复杂，实现高效清扫的关键在于如何做好路径规划。早期机器人缺乏路径规划能力，多采用随机碰撞法，即选定一个方向扫下去，发现障碍物后随机转向继续清扫。这种清扫方式的轨迹是凌乱的，而且有可能被困在某个位置无法脱身，如图1-35所示。

还有一种更有效的办法：把家里的地图告诉机器人，有了这幅地图，哪里有障碍，哪里有通路，就一目了然了。在此基础上，机器人就可以规划合理的清扫

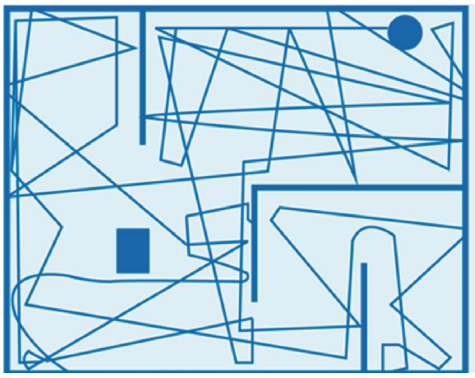


图1-35　采用随机碰撞法的扫地机器人的清扫轨迹

路径，既避开障碍，也可以防止重复清扫。这种方法显然比随机碰撞法更高效，但也存在一个关键缺陷，那就是不可能为每一台售出的扫地机器人都勘测一份家居地图；就算真能做到，当家里的家具位置发生变化后，原来的地图也会失效。为了解决这一问题，研究人员开发出一种同步地图定位与构造（SLAM）技术（图1-36），这种方法可以让机器边探索环境边把地图构造出来。有了SLAM技术，扫地机器人终于变得更加智能，能够有序清扫，而不是盲目乱撞了，如图1-37所示。

图1-36　SLAM算法示意图

注：粗线为机器人的运动轨迹，机器人一边运动一边构造环境地图。

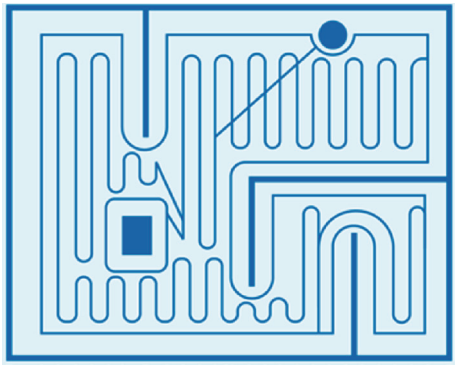


图1-37　具有路径规划能力的扫地机器人的清扫路径

② 自动驾驶汽车

自动驾驶指汽车能够在没有驾驶员直接操控的情况下自动行驶。汽车自动驾驶的探索可以追溯至20世纪20年代，当时称为“高级驾驶辅助系统”。1977年，日本科学家开始了半自动驾驶路上试验。2017年，Waymo公司宣布完成完全无人驾驶测试，次年在美国的凤凰城推出无人出租车服务。

现代自动驾驶系统依靠各种传感器（如激光雷达、摄像头、毫米波雷达等）感应周围环境，并通过人工智能算法来控制车辆行驶。其核心技术包括环境感知、路径规划和车辆控制。环境感知系统通过传感器实时获取周围环境信息。路径规划系统则根据当前环境和目标位置，生成最优行驶路线。车辆控制系统负责控制车辆，如加速、转向和制动。可以说，路径规划是自动驾驶的核心“智能”，其方法和扫地机器人相似，都基于SLAM算法。从路径规划角度看，自动驾驶汽车和扫地机器人有相似的算法基础，但自动驾驶汽车需要应对复杂的公路环境、更高的速度和更多不可预测的因素，其复杂程度远超扫地机器人。

值得说明的是，自动驾驶是一个统称，其自动化的程度也是分级的。按照国际汽车工程师学会的定义，自动驾驶分为5级，级别越高，自动化程度越高：第1级只是简单辅助功能，而第5级则完全不需要人类干预，能够在任何环境下进行自主驾驶。目前，很多公司正致力于实现第5级自动驾驶技术，并已经取得了不错的进展，如特斯拉已经推出没有方向盘和脚踏板的新车型，理论上完全不用人来控制。

自动驾驶会对未来的社会生活产生显著影响。首先，它将极大改变人类的出行方式，提高交通安全性和效率，减少交通事故和拥堵。此外，自动驾驶技术还将对物流运输、公共交通等领域产生深远影响。例如，自动驾驶卡车可以24小时不间断运行，大幅提高物流效率，降低运输成本；大型园区可以使用无人驾驶小巴不断穿梭运送乘客（图1-38）。未来，自动驾驶汽车将改变城市规划和基础设施建设，促进智慧城市的发展。

图1-38　北京首钢园里的“阿波龙”无人驾驶小巴

③ 机械手

不论是扫地机器人还是自动驾驶汽车都属于移动机器人，目的是实现高效、安全的位置移动。另一类是动作机器人，通过肢体动作完成特定任务，典型的是用于工业生产中的机械手。第一台工业机械手是由美国人乔治·德沃尔发明的Unimate（图1-39），后来应用于通用汽车的生产线。从此以后，工业机器人迅速发展，如今已经成为自动化工业流水线的核心工具。图1-40所示为一群机械手协同完成工业产品组装任务的场景。

图1-39　Unimate机器人为人冲咖啡

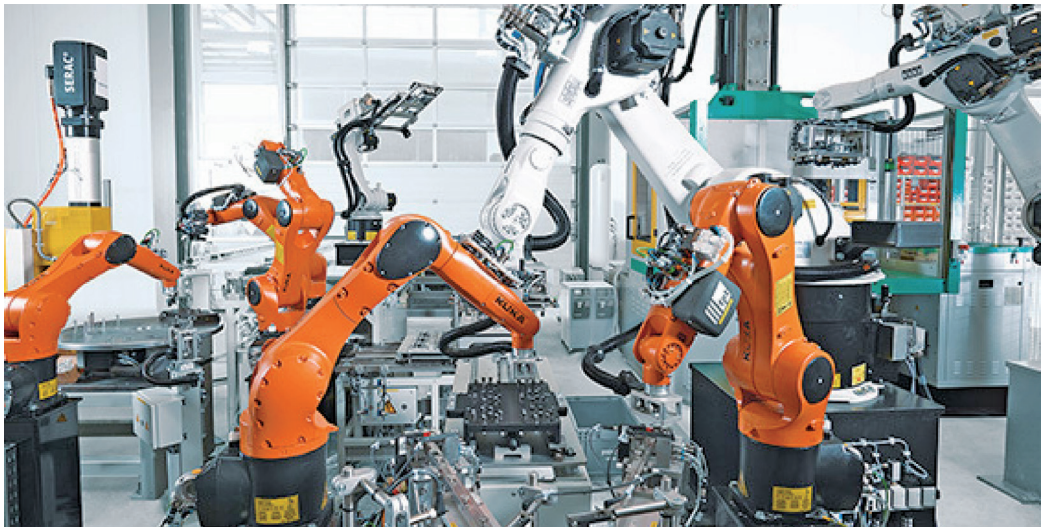


图1-40　现代工厂里协同工作的机械手

传统机械手的动作遵循人为设计的固定流程。需要事先严格设定各道工序，对误差的容忍度较低。比如，焊接件的位置出现偏差，机械手就可能无法定位焊点，从而影响任务完成。此外，人为设计的程序难以快速适应新任务，升级成本较高。随着技术的进步，越来越多的机械手引入了人工智能技术。比如通过示教的方式，牵引着机械手完成一次任务，它就可以学会如何操作，这种机器人称为示教机器人。更先进的机器人可以通过自我探索学会操作技能。比如美国加州大学伯克利分校设计的自动抓取机器人（图1-41），通过尝试各种抓取动作，能够自主学会抓取物品的技能。

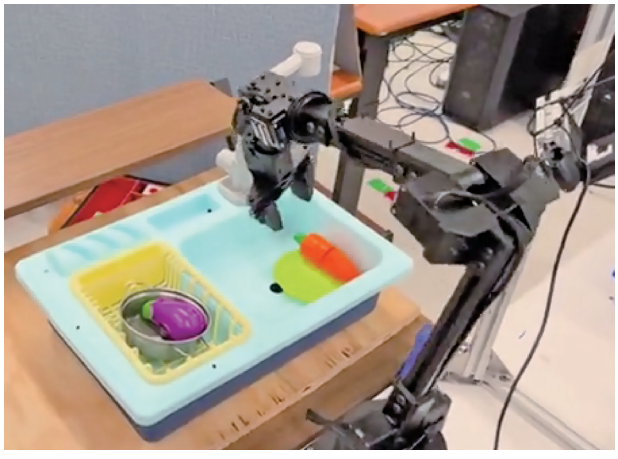


图1-41　美国加州大学伯克利分校设计的自动抓取机器人

工业机械手的广泛应用极大提高了生产效率和产品质量，降低了人力成本和劳动强度。例如，在汽车制造业中，机器人可以执行高精度的焊接和组装任务，大幅提高了生产线的效率和产品的一致性。此外，工业机器人在危险和高强度的工作环境中替代人类操作，提高了工作安全性。随着智能制造的普及，工业机械手将成为未来工厂的主力。

④ 总结

机器人是可以自主移动或执行动作的机器。并非所有机器人都具备智能，很多机器人执行的只是重复性动作。随着人工智能技术的进步，越来越多的机器人开始使用人工智能技术，从而变得越来越智能。例如，扫地机器人的移动本身并不智能，但具有规划能力的移动就具有了智能；对于工业机械手，按程序完成抓取动作智能性不高，但能手眼配合，可以抓取任意形状的物品就具有了极高的智能。未来，随着人工智能技术的不断进步，拥有了更高智能的机器人将在人类生活中扮演越来越重要的角色。