让机器听懂世界,触及人类梦想还有多远

陈孝良

中国科学院声学研究所,北京 100190

摘要 语言交互能力是人类认知发展、终身学习的基础,这为人类开启了智慧之门。人工智能时代,语言交互也将是人类和机器之间表达思想、交流知识、相互沟通的重要工具,这就需要让机器听懂复杂场景下的人类语言并且适应人类几千年进化形成的远场语音交互习惯,从而让机器真正认知人类世界,为机器产生类人智能提供一种参考。

关键词 麦克风阵列;语音识别;自然语言理解;远场语音交互

语言对于人类文明的重要性不言而喻,但是语言的起源却是个有高度争议的话题,我们对它几乎一无所知。我们虽然对动物、人类以及宇宙有了一些基本的认识,但对一些看似简单的问题远没有完全认识清楚,比如人类的耳朵为何要有这么奇怪的"耳廓"?

1 让机器听懂世界承载了人类千年梦想

语言承载了人类文化,人类需要通过语言学习知识和传递信息,这是人类区别于动物界最重要的特性之一[1]。人类语言超过了5000种,人类将大部分时间花费在学习各种语言上似乎不是一个更有效的途径。未来的机器智能时代,机器必然需要通过语言实现与人类之间的交互。因此,让机器听懂世界,这是未来机器智能时代的关键问题,也是人类一个更大的梦想,但是,我们距离人类的这个梦想还有多远呢?

让机器听懂世界,这其实蕴含了多个历程,包括听懂人类语言,进而听懂动物叫声,甚至听懂自然声音,亦或类似地球的耳朵LIGO那样聆听宇宙的"声音"。所有这些都是极其复杂的过程。人类没有这种能力,但是我们期望机器能够延伸人类的能力,从而实现人类的几千年的伟大梦想。

让机器听懂人类语言,这是最近技术和市场上非常火热的事情,也是全世界科学家为之奋斗了60多年的事业。这其中最为典型的,就是以亚马逊Echo所引领的智能音箱。迄今为止,占据全球市值排名榜的公司,包括中国的阿里巴巴、京东、腾讯、百度、小米等,国外的苹果、微软、亚马逊、谷歌、脸书、三星等,都在着力争夺未来智能时代的语音入口,甚至亚马逊和阿里率先不惜代价开启了补贴大战。这些全球巨头的激烈竞争,将对未来10年甚至20年语音智能的发展产生极其重要的影响。

那么,如何才能让机器听懂人类语言呢?这需要解决3个核心关键问题:听见、听准和听懂。从技术角度来看,就是拾音、识别和理解3个关键技术环节。拾音是最为基础的环节,必须保证让机器听得见声音,这部分主要是声学问题;识别是将符合要求的声音转化成文字,这部分主要是语音识别的问题;理解则是根据识别出来的文字,准确理解人类的指令甚或情感。鉴于语音智能设备已经大量出现在我们生活场景之中,当前技术的核心关键就是声学问题和语义理解[2-5]。

收稿日期:2017-11-22;修回日期:2018-02-03

作者简介:陈孝良,副研究员,研究方向为声视频融合,电子信箱:cxl@mail.ioa.ac.cn

引用格式:陈孝良. 让机器听懂世界,触及人类梦想还有多远[J]. 科技导报, 2018, 36(3): 36-40; doi: 10.3981/j.issn.1000-7857.2018.03.004

2　近场语音是机器听懂人类的率先尝试

近场语音交互主要是指人类距离机器不超过30 cm范围的语音识别技术,这项技术利用距离巧妙回避了真实场景下复杂的识别问题,可以理解为一种实验室理想环境下的语音交互技术。近场语音识别从20世纪50年代就开始研究,但是长期没有实质性进展,直到苹果公司在2010年推出Siri的应用,这才引起了全球的关注。到现在为止,近场语音交互技术已经比较成熟,平均识别率可以达到95%以上,主流的手机和平板等设备都已经普遍支持近场语音应用[2-7]。值得一提的是,很多人工智能大会或者电视演播厅所展示的实时语音识别或者翻译技术,其实都是近场语音交互技术,这些声音是从近场麦克风采集的高质量数据,与会场的嘈杂环境并没有实际关联。

但是近场语音交互受到了真实场景的巨大制约,并没有展现出来语音交互可以解放双手的先进性,因此在很多场景中,事实上近场语音交互都是"鸡肋"一般的存在,并没有发挥出真正的威力,也就说,这个技术其实被严重低估了。直到远场语音交互技术的出现成功解决了真实场景下的复杂声学问题以后,至少在技术上达到了用户认可的门槛,语音交互才真正出现了替代键盘、鼠标和触摸屏的可能性。

3　远场语音将语音智能落地到真实场景

远场语音交互主要解决0.3~5 m范围内的语音交互问题(图1),这个范围事实上就是人类之间沟通交流的最佳距离,距离太近容易触发自我保护意识,而距离太远则会增大交流难度[8]。注意语音交互并非只是语音问题,人类的交互其实是一个综合的过程,包括了表情、眼神、肢体动作等一系列影响因素,太远距离的语音交互事实上意义不大,比如隔墙的语音交互实际上只要做好语音控制就可以了,真实场景下并不需要复杂的交互设计,因为人类也很难在有隔挡的情况下愉快地对话。

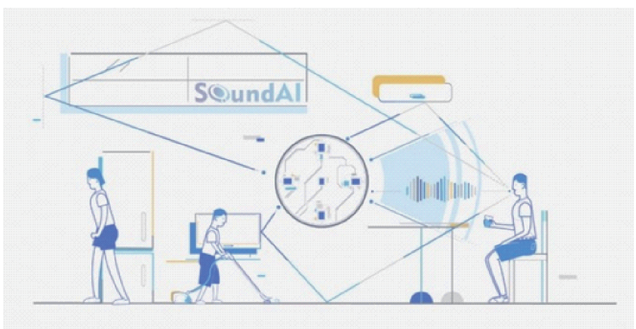

图1　复杂场景中的远场语音交互

远场语音交互的历史是比较短暂的,这项技术以前长期没有实质性突破。2014年是个重要的转折点,亚马逊的Echo开始探索这个市场,但是直到2016年末,全球才真正开始重视这项技术,并且短短1年时间,引领全球市场都进入了激烈博弈的阶段。目前,全球远场语音交互领域内的核心企业有亚马逊、谷歌、苹果、微软、百度、科大讯飞、声智科技、思必驰等。

当前,远场语音交互的代表产品自然是智能音箱,全球巨头都在智能音箱领域进行了重点布局。国际上各家巨头风格不一,亚马逊的Echo率先开启市场已然影响深远,谷歌的Home剑走偏锋以技术做博弈,微软的Invoke坚持工程师定义产品,苹果的HomePod由于低估技术难度导致无奈跳票,而脸书和三星则紧锣密鼓的研发追赶进度;反观中国国内更为热闹,小米的"小爱同学"以299元的低价撬开市场,阿里巴巴的"天猫精灵"则以99元的低价率先开启补贴,就在腾讯、华为还在犹豫的时候,百度刚刚发布了"渡鸦"智能音箱和DuerOS开发板SoundPi。虽然中国的智能音箱起步较晚,但是国内市场经常演绎奋起直追甚至超越的故事。

有两个重要的数据最具说服力:一个是亚马逊Echo的销量累计超过千万,2018年供应链下单已达到3千万台;另外一个数据就是阿里巴巴和小米的智能音箱在"双十一"活动中累计销量超过百万台,2018年预计销量超过千万台。这说明,智能音箱作为语音智能的突破口已经成立,这是远场语音交互的一大进步,再次说明只有落地真实场景并且经过验证的技术才具有生命力。这里要特别强调的是,智能音箱只是远场语音交互的突破口,并非语音的唯一入口,因为未来的机器智能时代,语音入口不仅仅只有智能音箱,其他设备比如电视、冰箱、汽车和机器人都有可能成为重要入口。但是智能音箱又是非常重要的,因为不管产品形态怎样变化,其本质其实还是智能音箱的基础技术架构。

4　让机器听懂世界有待解决的问题

若让机器听懂世界,远场语音交互技术也仅是个

尝试而已,事实上远场语音技术本身也只是刚刚起步,即便5 m以内,其噪声抑制、回声抵消、混响去除、远场唤醒和远场识别等核心技术还存在诸多缺陷[9-10]。但是,技术一直在迭代发展,特别是当技术落地实际场景以后,源源不断的真实数据和客户需求将带动技术更加快速的发展。

从技术层面来看,让机器听懂世界涉及了数学、物理学、语言学、医学、计算机学等各学科的知识,很难一一枚举出来,但是若从应用场景来看,则相对比较简单,让机器听懂世界包括了人类语言、人类情感、动物声音和自然声音。

4.1 听懂人类语言

近场和远场语音交互的技术可以解决5 m以内的语音交互问题,基本囊括了人机交互的主要问题,但是还有更多复杂场景的问题需要解决,如以下几个方面。

远场语音交互:主要解决5 m以内的唤醒、识别和理解问题,虽然这项技术已经落地实际的场景和产品,但是对于诸如鸡尾酒会效应(指人的听力选择能力,即注意力集中在某人的谈话之中而忽略背景中其他的对话或噪音)等难题仍然还没有实质性进展,而且从人类相互交流的过程来看,当前的远场语音交互技术还远远没有达到非常准确、非常顺畅的程度。

超远场交互:主要是指5 m、10 m、20 m甚至500 m以外的超远距离拾音和交互,这种技术的难度就是解决在远距离声音传播过程中能量衰减的约束下获取高质量声音数据的问题,因为没有高质量的声音数据,再厉害的机器学习也没有任何价值。这种技术主要应用在智能安防场景,例如交通监控,搭配远距离声发射技术可以实现远程指挥的自动交通处理。

局部场交互:主要是指针对某个局部范围内的语音识别和理解,主要应用于智能医疗、智慧法庭、智能教育、智能会议等特殊场景,例如实时记录和识别法官、医生或者教师说过的话。这种场景的需求比较单一,仅仅针对特定目标进行拾音和识别即可,但是对于识别的速度和精度要求非常高,一般要达到98%以上。

分布场交互:主要是指狭小空间内多人识别和响应的问题,最常见的就是汽车场景,现在的汽车智能交互仅仅照顾了驾驶员的需求,但实际应用中可能还需要照顾汽车上其他乘客的交互需求,这就涉及了多人识别和交互的问题。事实上,随着智能音箱等一系列智能设备的普及,未来我们的家庭就是典型的分布场交互场景。

多语种交互:主要适应跨语言时候的自由交互场景,当前Google、百度和科大讯飞推出的翻译机解决了部分问题,但是这些翻译机主要还是近场语音,过渡到远场语音交互的难度很大,因为翻译的场景确实太复杂多变了,在数据积累还没形成规模之前,这类技术还很难有实质性突破。

大词汇交互:语音识别能否应用到话剧的场景,这似乎是一个更加令人头疼的问题,因为无论从声学、识别还是到理解都是巨大的挑战。话剧演员一般不会佩戴麦克风,这就要求远场多人识别,而且话剧演员常会自白一大段,如何进行端点识别和语音识别?这样发散想来,当前的智能语音技术真的是才刚刚开始。

4.2 听懂人类情感

至于听懂人类情感,则是一个更加复杂的过程。至今,人类自己也没搞清楚情感的来源,例如即便热恋中的情侣,也无法搞清楚对方的真实需求。但是至少有几个技术点是和人类情感有关系的。

声纹识别:它的理论基础是每一个声音都具有独特的特征,通过该特征能将不同人的声音进行有效的区分。声纹的特征主要由2个因素决定,一是声腔的尺寸,具体包括咽喉、鼻腔和口腔等,这些器官的形状、尺寸和位置决定了声带张力的大小和声音频率的范围;二是发声器官被操纵的方式,发声器官包括唇、齿、舌、软腭及腭肌肉等,他们之间相互作用就会产生清晰的语音,而它们之间的协作方式是人通过后天与周围人的交流中随机学习到的。

情感识别:主要是从采集到的语音信号中提取表达情感的声学特征,并找出这些声学特征与人类情感的映射关系。情感识别当前也主要采用深度学习的方法,这就需要建立对情感空间的描述以及形成足够多的情感语料库。情感识别是人机交互中体现智能的应用,但是到目前为止,技术水平还没有达到产品应用的程度。

哼唱识别:主要是通过用户哼唱歌曲的曲调,然后通过其中的旋律同音乐库中的数据进行详细分析和比对,最后将符合这个旋律的歌曲信息提供给用户。目前这项技术在音乐搜索中已经使用,识别率可以达到80%左右。

声光融合:声学和光学总是相伴相生,人类的情感也是通过听觉和视觉同时接受分析的,因此机器将语

音和图像结合在一起分析,才能更好地理解人类的情感,但是语音和图像在各自领域并没有发展成熟,因此声光融合的研究一直处于被轻视的尴尬地位。

4.3　听懂动物声音

让机器听懂动物的声音,或许是一个苛刻的要求,因为人类至今也没有听懂动物的声音。甚至是婴儿的哭声,我们也只能大概猜测。但是这不影响机器的进步,因为在很多领域,机器迟早是要超越人类的。事实上,这类研究一直在进行,比如海豚、蝙蝠、鲸鱼、猩猩、老虎、狮子、猫、狗、蚊子、蜂鸟等动物的声音特征,当数据积累足够多的时候,根据声音推断这些动物的行为不是不可能,而人类的进步很大程度也得益于这种仿生。

4.4　听懂自然声音

当然,机器也必须听懂大自然的声音,例如雷声、雨声、地震、海浪、风声等,通过这些声音则可以辨别机器所处的环境,并且根据环境做出判断。这些技术也正在产业化落地,例如声智科技与百度联合研究的小样本学习技术,其中一个特性就是可以根据噪声来判断场景变化,显然厨房和客厅、卧室的噪声不会相同,同样地,咖啡厅、火车站、机场、办公室、汽车等场景的噪声也有很大区别,通过区分这些噪音则可以快速匹配出场景,这将非常有利于后端智能的处理,对自然语言的理解会更加准确。

5　听懂世界还需要更多硬科技的尝试

目前来看,语音交互的精度和速度取决于实际应用环境,学术界探讨了很多语音交互的技术趋势,有两个思路是非常值得关注的,一个是端到端的机器学习系统,另一个是G.E. Hinton最近提出的胶囊理论。胶囊理论在学术上争议比较大,能否在语音交互领域体现出来优势还值得探讨[11]。

但是,让机器听懂世界,不能仅仅依赖算法和数据,更重要的还是底层硬科技的突破[12-15],如一些有关传感器件的基础技术。

智能麦克风:可以简单理解为将当前的MEMS麦克风与低功耗芯片融合在一起,主要是解决低功耗语音唤醒和识别的问题。

矢量麦克风:当前的麦克风都是标量麦克风,只能获取单一的物理信息,也就是能量值,根据时间信息和阵列配置才能获取频域和相位信息。若将标量麦克风升级成为矢量麦克风,则增加了一个维度的特征信息,这对于机器学习的提升将可能会非常明显。

薄膜麦克风:这是一种柔性的技术,可以想象把整个电视屏幕当作麦克风的场景,通过特殊的纳米材料技术,甚至可以把任何界面都当作声音的接收装置,通常来说这种换能器装置也可以把声音转变成电能。

柔性扬声器:这实际上和薄膜麦克风的原理类似,只是将换能的方向换了一下。柔性扬声器目前有多种方案,目前来看,其难点主要还是发声的带宽和失真问题。

激光拾声:这是主动拾声的一种方式,可以通过激光的反射等方法拾取远处的振动信息,从而还原成为声音,这种方法以前主要应用在窃听和安全领域,目前来看,由于还原声音的质量还不够,这种方法应用到语音识别还比较困难。

微波拾声:微波是指波长介于红外线和无线电波之间的电磁波,频率范围大约在300 MHz~300 GHz,它同激光拾声的原理类似,只是微波对于玻璃、塑料和瓷器几乎是穿越而不被吸收。

高速摄像头拾声:这是利用高速摄像机来拾取振动从而还原声音,这种方式需要可视范围和高速摄像机,只在一些特定场景里面应用。

6　结论

让机器听懂世界的技术正在全球范围内快速演化,相信不久的将来,肯定能看到更加智能的机器,因此,既不要抨击当前的人工智能技术,也不要盛赞现在的基础科学技术,保持一颗平静的心,正确给予科技界和产业界支持才是对未来人工智能技术最大的支持。一项技术的价值最终会体现在它为社会创造的价值上面。

但应该看到,中国产业界长期不重视基础技术研发的投入,科研院所的经费严重不足,甚至资本界也常常不看好技术类型公司,这实际上严重束缚了相关技术的发展,这和美国形成很大反差。当然,这也是经济发展的必经阶段,只有解决了经济问题,我们才能真正对于知识产生自由的渴望,才能看的更远,才能追求更大的梦想。

参考文献(References)

[1] Jackson H, Stockwell P. An introduction to the nature and functions of language[M]. New York & London: Continuum International Publishing Group, 2010.

[2] Jurafsky D, Martin J H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition[J]. 2000, 36(23): 161-187.

[3] Keshet J, Bengio S. Automatic speech and speaker recognition: Large margin and kernel methods[M]. West Susse: Wiley, 2009.

[4] Huang X, Acero A, Hon H W. Spoken language processing: A guide to theory, algorithm, and system development[M]. New Jersey: Prentice Hall, 2001.

[5] Rabiner L, Juang B H. Fundamentals of speech recognition[M]. Beijing: Tsinghua University Press, 1999.

[6] Jurafsky D, Martin J H. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition[J]. 2000, 36(23): 161-187.

[7] Li D, Dong Y. Deep learning: Methods and applications[J]. Foundations and Trends® in sigal processing, 2014,7(3-4).

[8] Angus J, Howard D. Acoustics and Psychoacoustics, 3rd edition [J]. Elsevier Ltd Oxford, 2016, 54: 365-436.

[9] 程建春. 声学原理[M]. 北京: 科学出版社, 2012.
Cheng Jianchun. Acoustics principle[M]. Beijing: Science Press, 2012.

[10] Everest F A, Pohlmann K C. Master handbook of acoustics [M]. New York: McGraw-Hill, 2001.

[11] Sabour S, Frosst N, Hinton G E. Dynamic routing between capsules[J]. Neural Information Processing Systems, 2017.

[12] Beranek L L, Mellow T J. Acoustics: Sound fields and transducers[M]. Oxford& Waltham: Elsevier, 2012: 449-479.

[13] Ma G, Yang M, Sheng P, et al. Acoustic metamaterial with simultaneously negative effective mass density and bulk modulus: US, US 8857564 B2[P]. 2014.

[14] Greif S, Zsebök S, Schmieder D, et al. Acoustic mirrors as sensory traps for bats[J]. Science, 2017, 357(6355): 1045.

[15] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

Making the machine understand the human world

CHEN Xiaoliang

Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China

Abstract The ability of language is a basis of human cognitive development and lifelong learning, which opens the door for human wisdom. In the era of artificial intelligence, language is also an indispensible tool for the machine to express ideas, exchange knowledge and communicate with human world. The key to make the machine truly recognize the human world is to let the machine not only understand human language in complex scenarios but also adapt to the far-field voice interaction habits that have been formed by human evolution for thousands of years. This article hopes to provide a reference for development of machines with human intelligence.

Keywords microphone array; automation speech recognition; natural language processing; far-field speech interaction

(责任编辑 王丽娜)