中国科学: 物理学 力学 天文学 2022 年 第 52 卷 第 4 期: 244305
SCIENTIA SINICA Physica, Mechanica & Astronomica physcn.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS
进 展 声学科学与技术前沿论坛专题

语言声学智能化的思考与探索

颜永红[1,2*], 程高峰[1,2]

1. 中国科学院声学研究所, 北京 100190;
2. 中国科学院大学, 北京 100049
*联系人, E-mail: yanyonghong@hccl.ioa.ac.cn

收稿日期: 2021-11-03; 接受日期: 2022-03-09; 网络出版日期: 2022-03-22
北京市科学技术委员会(编号: Z211100002521020)资助项目

摘要 人工智能的热潮正在席卷各行各业, 声学领域也在进行声学智能化的探索, 即如何实现声学技术与人工智能技术的有机融合和应用. 本文着重以语言声学为例, 同时结合医疗声学, 首先分析声学智能化所面临的数据稀缺性、算法自主性、算力依赖性、系统安全性以及人才短缺等重大挑战, 进而对与之相关的关键技术进行总结, 探讨可能的解决方案. 在此基础之上, 回顾近年来国内外智能声学方面的主要研究进展, 并总结国内智能语音的产业化应用发展历程. 最后对声学智能化的发展方向和趋势做出展望, 针对智能语音和医疗声学方面的具体问题提出发展建议.

关键词 语言声学, 声学智能化, 医疗声学

PACS: 43.60.Np, 43.71.Sy, 43.72.Bs, 43.72.Ne, 43.80.Qf

1 引言

大力发展人工智能是当前的重要国家战略. 2017年《新一代人工智能发展规划》指出: “类脑智能计算理论重点突破类脑的信息编码、处理、记忆、学习与推理理论, 形成类脑复杂系统及类脑控制等理论与方法, 建立大规模类脑智能计算的新模型和脑启发的认知计算模型.” “跨媒体感知计算理论重点突破低成本低能耗智能感知、复杂场景主动感知、自然环境听觉与言语感知、多媒体自主学习等理论方法, 实现超人感知和高动态、高维度、多模式分布式大场景感知.”

我国目前在布局“建立新一代人工智能基础理论体系”, 并在此基础上, 建立新一代人工智能关键共性技术体系.

人工智能包含三要素: 数据、算力、算法; 建模方式包括卷积神经网络(CNN)[1−3]、循环神经网络(RNN)[1,4]、长短时记忆深度神经网络(LSTM)[5]、双向长短时记忆网络(BLSTM)[6]等; 学习方式包括迁移学习[7]、增强学习[8]、对抗学习[9]等.

声音是人类感知环境、交流沟通的主要方式之一, 其中包含大量信息, 人们一直在探究利用声学技术来高效获取信息、提升效率的方式. 作为涉及诸多领域的交叉学科, 近几十年来声学学科在各专业研究人员的不懈努力下不断取得突破. 在当今人工智能的发

引用格式: 颜永红, 程高峰. 语言声学智能化的思考与探索. 中国科学: 物理学 力学 天文学, 2022, 52: 244305
Yan Y H, Cheng G F. Thinking and exploration into intellectualization of speech and medical acoustics (in Chinese). Sci Sin-Phys Mech Astron, 2022, 52: 244305, doi: 10.1360/SSPMA-2021-0306

展浪潮中, 如何实现声学智能化, 推动声学技术与人工智能技术的有机融合和应用, 是各高校、科研院所、企业等一直致力探索的前沿方向. 下面以语言声学和医疗声学的智能化应用为例简要介绍.

语言声学主要研究语音的发音过程和机理、语言的物理性质、语言的感知、语言的处理、语音的通信、机器自动识别和理解语言以及语音和文字相互转换等问题. 语言声学智能化涉及的核心技术包括: 音频信号采集和处理、传声器阵列处理、语音增强、语音识别、语音合成、语义理解和音频场景分析等等. 语言声学和人工智能结合发展而来的人机自然语音交互技术是人机交互最方便快捷的手段, 语音正在成为移动互联网的最重要入口, 以语音交互为基础的人机互动模式正在逐步成为手机、汽车、智能家居、智能家电、智能玩具等相关产品的标准配置. 此外, 由于传统客服业务人工成本居高不下, 基于语音能力的智能客服机器人、客服质检系统等的需求方兴未艾. 其他的应用场景还包括: 会议语音自动转写记录、法院智能语音庭审、广播电视网络多媒体音频自动分析等.

医学声学, 即通过对声音相关信息的自动特征提取和分析, 进行人体疾病的自动筛查检测和诊断, 从而大幅度降低医疗成本, 节约医疗资源. 通过引入人工智能技术, 目前国际上医疗领域内已有多项人类疾病可以进行自动检测. 该领域涉及的核心技术包括: 心肺音/咳嗽声等信号采集、处理和降噪、特征提取、自动分类和疾病诊断等. 在医学声学方面, 以心血管病的检测为例: 心血管疾病是一种容易导致人类死亡的疾病, 对其的监测和诊断, 对改善人类健康状况意义重大. 而传统的诊断方法所用设备价格较高、且需要专人操作, 不易推广. 由于心音是心脏及心血管系统机械运动状况的反映, 心音信号的分析与识别对心血管系统疾病的诊断筛查有着重要的潜在应用价值, 因此声学技术与人工智能的结合具有重要的研究价值和发展前景.

经过多年发展, 当前国内语音信息处理与医疗声学等技术已基本与国外主流研究机构比肩, 多家科研单位都取得了丰硕成果, 国内语音技术研发机构在领域内的国内、国际评测中成绩优异. 但相关声学技术的改进和应用仍然面临多重困难, 对我国相关领域的研发人员具有更高要求与挑战.

本文主要内容安排如下: 第2节介绍智能语音领域的主要科学问题和关键技术; 第3节回顾国内外语言声学智能化近年来的主要研发进展, 简要介绍智能医疗声学的应用研究; 第4节为对未来发展方向、趋势的展望和建议.

2 智能语音挑战与关键技术

2.1 智能语音挑战

声学智能化依托于人工智能技术的发展, 也受到人工智能要素的制约. 本节首先探讨智能语音所面临的如下挑战.

(1) 数据的稀缺性问题

高价值目标的数据总是稀缺的, 数据采集困难、采集成本高、标注成本高. 如何克服主流有监督方法对高成本标注数据的过度依赖, 发展适用于低资源场景下人机交互的语言声学技术, 是重要的科学问题.

(2) 对开源算法的依赖

当前业界的人工智能模型训练平台和算法, 多为国外单位研发, 例如: 英国剑桥大学的HTK[10], 美国约翰霍普金斯大学的Kaldi[11], 日本Shinji Watanabe教授研究小组[12]主导的ESPNet等. 当前国内的人工智能技术研发仍然高度依赖国外平台, 存在巨大的卡脖子风险. 因此我国迫切需要研发具有自主知识产权人工智能工具平台和算法, 改变对国外相关技术平台高度依赖的局面, 实现声学智能化技术的安全可控.

(3) 复杂场景下鲁棒语音信息处理问题

(i) 如何实现复杂场景下的语音分离与识别?

在真实场景中, 由于存在噪声、混响和多说话人等各种干扰, 智能语音处理的性能会受到严重影响. 人类在分离语音时, 采取渐进式的解决方式, 即识别说话人、分离说话人和理解分离语音. 依靠该机制, 听觉系统能够自动调整其注意力, 从而拾取目标说话人信号, 而后通过神经系统实现对信号的理解. 因此, 如何结合人脑听感知机制和信号模型, 探索复杂场景下基于深度学习的语音分离与识别统一框架, 是当前迫切需要研究的一个关键科学问题.

(ii) 如何解决语音识别和声纹识别中的跨信道问题?

语音识别和声纹识别在跨信道条件下性能受到严重干扰, 这也是我们当前需要重点解决的问题. 首先是信道不匹配对识别系统性能的影响, 例如某应用需求

涉及特定通信信道, 而常用声纹系统的背景模型往往由电话信道、互联网信道等语音数据训练得到, 信道变化会造成巨大差异. 信道不仅种类繁多、特性未知, 且信道特性往往与语音特性绑定紧密, 难以充分分离和挖掘. 如何准确去除语音中的信道干扰信息, 从而大幅减弱其对语音识别和声纹识别性能的影响, 是当前的一个研究热点.

(4) 对算力的依赖

成本是技术无所不在的主要制约之一. 目前的人工智能算法对于算力要求很高, 模型训练一般需要众多的服务器资源和GPU卡资源, 硬件投入十分巨大. 如何降低算法对算力的依赖, 减少声学智能化的成本投入, 也是重要的技术问题.

(5) 人才稀缺问题

目前声学智能化领域内会用工具的人员多, 研究原理、掌握核心算法的人员少. 迫切需要将研究成果工具化, 易于一般工作人员使用. 同时瞄准典型应用, 加强产学研用的密切合作, 以便技术的推广和应用.

(6) 智能语音发展中的安全问题

随着人工智能和深度学习算法的高速发展, 这些新兴技术在音视频识别、自然语言处理等领域已经得到了广泛应用. 然而, 当前主流人工智能模型的安全性是语言声学智能化过程中无法忽视的问题.

端到端深度学习模型是目前语音技术的主流, 在语音识别和语音合成等方面处于领先地位. 尽管它们在许多应用中都有出色的性能精度, 但也很容易被恶意的攻击者所欺骗, 强迫模型产生错误结果, 甚至产生一个目标输出值. 这种对抗性样本攻击已被证明对图像识别、物体检测以及语音识别非常成功. 针对语音对抗项目, 在获取目标说话人的数据后, 经过数据处理, 通过神经网络的训练, 可以在保留源说话人的内容信息的同时, 具有目标说话人的音色信息, 从而生成伪造语音. 在语音识别领域, 对抗样本生成技术研究进展迅速, 针对固定模型结构、固定模型参数已经可以对语音识别模型展开有效的攻击, 在听感无明显差别情况下对语音识别转写结果产生定向或非定向的干扰攻击. 随着智能交互设备的推广使用, 语音已经成为人机交互的重要接口, 针对对抗样本攻击进行有效的防御性研究, 提升语音系统的安全性是当前一个重要的研究方向.

隐私信息安全性问题同样需要重视. 训练数据隐私窃取主要包括数据还原、成员推理、属性推理等, 不需要直接接触训练数据, 攻击者即可根据模型输出等信息, 反向推断隐私数据、判断特定数据存在与否、推断隐私数据类型. 该类数据窃取攻击仅通过云服务交互接口即可进行, 方便实施, 成本较低, 却能造成严重的隐私泄露问题. 此外, 目前大多数机器学习任务在训练和预测过程中资源需求巨大, 计算及存储成本高昂, 模型的参数需要得到保护. 目前的攻击方法可以分为两步: 1) 向目标模型查询一组输入语音, 并获得模型给出的预测; 2) 使用上一步得到的“语音-预测”对训练一个仿制品. 通过进行攻击, 可以得到一个仿制品模型, 其功能与目标模型相近, 却不需要训练目标模型所需的金钱、时间、脑力劳动的开销, 造成模型开发者的损失.

2.2 智能语音关键技术

针对声学智能化领域面临的重重困难, 本节对该领域的一些技术进展进行总结.

2.2.1 利用无标注数据和低资源语音建模技术, 减少声学智能系统对数据的依赖

(1) 无监督语音检测技术

针对复杂环境下自然口语的语音感知问题, 研究非线性增强处理对语音感知的影响, 采用基于无监督学习框架的语音活动度检测统计模型. 该框架可由序贯高斯混合模型实现[13], 由于使用了无监督学习策略, 该语音活动度检测方法不需要大多数同类方法所依赖的常见基本假设, 即音频的前几帧是非语音的. 这一特点使该方法具有更好的普适性和容错性, 从而能够对复杂环境的语音进行更为有效的分段和检测.

(2) 基于自监督学习的语音预训练技术

在现实场景下, 有标注语音数据的数据量通常是十分有限的, 同时, 无标注的语音数据则可以通过相对低的成本收集得到. 因此, 利用无标注数据通过自监督学习的方法对语音识别模型进行预训练[14], 是使少量标注数据达到海量标注训练数据效果的关键技术. 自监督学习方法针对语音识别的具体应用可以分为声学预训练[15–18]和语言学预训练[19–21]. 前者利用大量的无标注语音学习更适用于语音识别任务的语音特征表示, 学习到的预训练模型可以用来做语音识别模型的特征提取器, 也可以用来做端到端模型的编码器

(Encoder)的初始化. 而后者则利用大量的文本对端到端模型的解码器(Decoder)进行无监督预训练, 帮助模型更好地进行语言学层面的建模.

(3) 基于半监督学习的语音识别技术

自监督学习方法以随机初始化的模型为起点, 依靠精心设计的损失函数让模型从语音本身学习到一种优化的特征表示; 而半监督学习则以经过有监督训练的模型为起点, 利用额外的无标注数据来优化模型[22]. 该方案与自监督学习分别在模型的不同阶段起到利用无标注数据的作用, 二者可以很好地融合在一个系统当中[23]. 具体来说, 半监督学习方法是将有限的人工标注音频、大量的无标注音频和纯文本语料结合, 从而实现与大量人工标注音频训练的语音识别系统相同的性能.

(4) 基于迁移学习的语音识别知识共享技术

利用具有丰富资源的多种语言的语音识别声学建模数据, 进行基于迁移学习的多语种声学建模, 是解决小语种数据资源受限的一种途径[24,25]. 迁移学习可以通过借鉴丰富资源领域的知识来提升资源匮乏领域的模型性能, 是机器学习领域内广为使用的方法. 在语音识别中, 可以利用迁移学习来进行多语种声学建模, 通过参数共享的方式利用其他语种的标注数据来辅助目标语种的语音识别模型训练[26,27].

2.2.2 着眼复杂音频场景, 研发鲁棒智能声学处理技术

(1) 复杂声学场景下融合信号处理和深度学习的语音拾取技术

研究不同应用场景下基于深度神经网络的语音降噪算法. 在基于深度神经网络的单通道降噪算法中, 结合体现人耳感知特性的频带权重函数, 建立模型优化准则, 进一步提升降噪后语音的感知可懂度[28,29]; 针对多通道语音降噪问题, 利用基于深度神经网络的单通道掩蔽估计方法, 提高波束形成方法中的语音协方差估计准确度, 提升降噪后语音的识别率表现[30]; 在双耳语音降噪任务中, 结合双耳空间特征, 对双耳信号的复数掩蔽进行联合估计, 去除噪声并保持目标信号的空间信息[31].

(2) 复杂场景下有效的语音表达特征提取技术

分别从滤波器感受视野和函数表达两个方面, 从时域和频域两个角度, 实现语音滤波器的设计. 其中感受视野可以建模成窗长等参数, 函数表达可以建模成滤波器形状等参数, 再分别利用完备频域和过完备时域基底, 构建特定场景下采集的语音信号的表达. 研究基于多尺度小波特征的音频场景识别技术, 提出适用于场景音频表征的多尺度小波特征, 构建完全基于深度学习的迭代式数据扩增框架, 结合自适应滤波卷积分类器, 提升音频场景识别准确率[32–34].

(3) 基于对抗性学习的领域自适应技术

由于经过不同信道传输的语音数据的声学差异性, 根据一种信道的语音数据训练的识别系统, 通常对其他信道传输的语音识别准确率较低. 作为语音可变性的主要来源之一, 信道对语音识别的鲁棒性提出了巨大的技术挑战. 跨领域建模的技术路线是: 以大量的源域数据训练生成对抗式网络, 训练同时具有自动语音识别系统鉴别性和跨领域不变性的中间层特征. 再将中间层特征输入到下游语音识别系统, 以实现对不同信道的语音信号鲁棒的端到端自动语音识别系统[35,36].

2.2.3 采用非自回归模块化端到端和流式端到端语音识别技术

自回归端对端系统拟合能力强, 但由于其自回归特性导致系统在GPU上加速效果不理想. 采用基于非自回归的模块化端对端识别技术[37–39], 相比于传统自回归端对端系统能够取得显著加速效果. 注意力机制[40–42]是当前端到端语音识别被广泛研究的架构之一. 目前基于注意力机制架构的语音转写系统虽然能达到较高的准确率, 但是其推理时需要整段语音的输入, 无法被用于实时识别的任务. 采用流式的注意力机制[43–45]对于实时场景下的语音识别意义巨大, 能够提升解码速度和效果, 减少算力依赖.

2.2.4 语音对话关键技术

智能语音对话系统(Spoken Dialog System)主要分为开放域对话系统和任务导向对话系统两类, 包含5个技术模块: 语音识别(Automatic Speech Recognition, ASR)、自然语言理解(Natural Language Understanding, NLU)、对话管理(Dialog Management, DM)、自然语言生成(Natural Language Generation, NLG)和语音合成(Text to Speech, TTS). 其中语音识别与语音合成是语言声学智能化相关的主要技术.

(1) 语音识别技术

将语音流识别为对应文本是对话式交互流程的第

一部分, 随着深度学习的广泛应用, 端到端语音识别模型逐渐成为主流, 注意力机制[40–42]是当前常用的架构之一. 自监督学习[14]和半监督学习技术[22]的兴起进一步推动了语音识别技术的发展. 基于自监督学习的声学预训练[15–18]依靠精心设计的损失函数, 让随机初始化的模型从语音数据本身学习适用于语音识别任务的特征表示. 声学预训练模型可以作为特征提取器或用来初始化端到端模型的编码器. 半监督学习以有监督训练的模型为起点, 将有限的人工标注音频、大量的无标注音频和纯文本语料结合, 从而实现与大量人工标注音频训练的语音识别系统相同甚至更优的性能. 二者融合在一个系统中[23]有效利用了低成本的无标注数据, 降低了对标注数据的依赖.

(2) 语音合成技术

将对话系统NLG部分输出的文本形式的内容转换为高可懂度和自然度的语音反馈给用户, 是实现人机沟通的关键技术. 近年来, 随着深度学习的发展, 基于神经网络的语音合成极大地提高了合成语音的质量. 经典的语音合成包括文本分析、声学模型、声码器三个模块, 实现文本字符、音素或语言学特征、声学特征以及语音波形之间的转换. 目前一些完全端到端的TTS模型则直接实现文本到语音波形的合成. Char2-Wav[46]利用基于RNN的Encoder-Attention-Decoder模型由字符生成声学特征, 然后使用SampleRNN[47]生成语音波形, 两个模型联合训练用于直接语音合成. 与之类似, ClariNet[48]联合训练自回归声学模型和非自回归声码器以直接生成波形. FastSpeech 2s[49]用完全并行的结构直接从文本生成语音, 大大加速了推理过程, 它利用一个辅助的mel谱图解码器帮助学习音素序列的上下文表示, 以减轻文本到波形联合训练的困难. 并行的EATS[50]利用持续时间插值和软动态时间弯折损失进行端到端学习, 实现和输入对齐, 同样直接从字符或音素生成波形. Wave Tacotron[51]在Tacotron的基础上构建了一个基于流式的解码器直接生成波形, 在该部分采用并行波形生成, 但在Tacotron部分仍然使用自回归生成方式.

3 国内外近年声学技术进展

3.1 国外研发进展

声学智能是当前人工智能的应用热点之一. 近年来随着深度学习技术的推动, 智能声学技术呈爆发式增长态势, 网络结构也越来越复杂, 如卷积神经网络(CNN)[1–3]、循环神经网络(RNN)[1,4]、长短时记忆网络(LSTM)[5]、双向长短时记忆网络(BLSTM)[6]、延时神经网络(TDNN)[52,53]、残差网络(Residual Network)[54]等相继被应用于声学模型建模和语言模型建模, 不断刷新着语音识别系统的精度. 基于深度神经网络技术, 各大公司纷纷发布了自己的智能语音商用系统, 如微软的小娜、小冰, 苹果的siri, 谷歌的google now, 亚马逊echo音箱和MIT的伴读机器人Jibo等.

谷歌推出了几款开源软件: Dialogflow是一款聊天机器人, 利用了语义理解技术; Duplex是针对特定场景的语音助手, 帮助用户预约餐厅、发廊等; Contact Center AI是一款语音导航+客服助手, 可以替代人工客服接听电话. Contact Center AI的客服助手功能支持利用语义理解和语音识别技术进行在线质检和话术推送, 对坐席进行监控、提醒和推送, 但目前使用的效果不是十分理想, 实时性上存在一些明显延迟, 并且提醒和推送的内容有时会不准确. 微软公司推出的新一代微软“小冰”已经实现了全双工语音对话的部署. Facebook提供Messenger Broadcasts (广播), 为企业用户提供一系列自动化客服支持和互动体验, 通过大流量赋能小微企业, 提供结合广告营销和智能客服的全线服务.

美国麻省理工学院为代表的欧美国家研究机构对利用咳嗽声进行新冠肺炎诊断的方法进行了探索. 选取的声音相关特征包括: Muscular Degradation (肌肉疲劳), Vocal Cords (声带变化), Sentiment/Mood (认知衰退、怀疑、挫折感). 美国麻省理工学院的研究结果表明, 利用咳嗽声可高精度发现无症状患者, 为低成本主动发现提供了一条新途径. 目前已和多家医院合作进行临床实验.

3.2 国内研究进展

国内语音与声学领域的主要研究单位包括中国科学院声学研究所(中科院声学所)、科大讯飞、清华大学、北京大学、中国科学院自动化研究所、百度、阿里、腾讯等研究所、高校和企业.

中科院声学所提出了Per-Frame Dropout (帧级别丢弃算法)[55], 通过帧级别的随机扰动阻断固定历史轨迹的形成, 解决了序列化训练下, LSTM对不同长度语

音输入的不鲁棒性问题; 提出了SOC (半正交限定技术)技术[56], 可以降低神经网络的优化难度, 增强神经网络的收敛效果; 提出了OPGRU (带输出门的低维映射门控递归神经网络单元)技术[57], 计算复杂度更低、收敛速度更快, 能够更好利用深度神经网络上下文信息进行声学建模. 上述成果都已被国际主流语音识别开源软件kaldi采纳, 其中基于SOC技术的TDNN-F模型还被多支参加语音界著名的多通道语音识别比赛Chime-5的队伍采用. 中科院声学所还提出业界首套流式端对端联合解码解决方案, 显著提升了在线语音识别系统的性能, 包括sMoChA注意力机制[45]、单调截断注意力机制(MTA)[58]、动态等待联合解码(Dynamic Waiting Joint Decoding)机制[45]对齐局部编码器和注意力机制的解码过程, 使得二者能够联合用于流式解码等领域. 中科院声学所还研发了新一代安全可控的ThinkITSpeechPlatform语音模型训练平台, 摆脱了对国外核心技术平台的依赖. 科大讯飞提出了前馈型序列记忆网络FSMN (Feed-forward Sequential Memory Network)[59]的语音识别建模方法, 改进了RNN中的梯度消失问题, 使其具有类似LSTM的长时记忆能力. 阿里巴巴提出Deep-FSMN (DFSMN)[60]的网络结构, 在训练过程中, 高层记忆模块的梯度会直接赋值给低层的记忆模块, 从而改进网络的深度造成的梯度消失问题. 百度提出了基于复数CNN网络的语音增强算法[61].

在产业化应用方面, 智能语音在众多领域都有巨大需求, 例如金融、电商、教育、医疗、生活服务等各个行业. 市场需求按年增长迅速, 2020年中国智能语音市场规模达到113.96亿元, 同比增长19.2%, 预计2026年中国智能语音市场规模将达到326.88亿元. 我国智能语音从业人员稀缺, 人工智能人才不足两万, 杰出人才不足一千. 其中, 智能语音相关人才更为稀缺.

智能声学的技术也应用在医疗方面, 中科院声学所实现了对于心音和肺音的特征提取和相关疾病的初步分析诊断. 提出了基于BiLSTM的心音定位算法对听诊位置进行定位, 融合多区域结果进行联合诊断, 对心脏疾病进行严重等级分类(轻、中、高). 提出采用IVA语音分离算法对心音和肺音进行分离, 减少心音对肺音的干扰; 采用OMLSA肺音降噪算法, 降低噪声对识别结果的影响; 研究了生成模型的肺音分类算法, 实现对常见肺部疾病初级诊断(正常肺音、啰音、哮鸣音、胸摩擦音), 等等.

4 声学技术智能化展望

结合国内外研究进展, 我们认为, 声学技术与人工智能的结合具有重要的研究价值和发展前景. 针对未来的发展方向与趋势, 本文提出如下建议.

(1) 发展小资源、无监督条件下的多语言语音建模方法, 减少对数据的依赖.

(i) 发展无监督学习技术, 利用大量未标注的语音数据, 实现建模和优化

在语音识别的无监督学习方面, 如何有效地在无监督条件下对语音序列建模; 语音具有复杂的层次结构(音素、音节、单词等), 如何在无监督的条件下学习这些结构性信息, 这些问题都很具有挑战. 现有的无监督语音识别模型收敛过慢, 通常需要消耗大量的计算资源和时间成本来完成, 因此需要研究新方法来加快无监督模型的收敛速度, 降低训练成本. 目前无监督的语音识别方法大多基于联合预训练声学模型和语言模型, 未来的研究方法需要将声学模型和语言模型进行解耦预训练, 以简化训练流程.

(ii) 发展迁移学习技术, 实现新语种的建模和识别

在语音识别的迁移学习方面, 对于丰富资源语种向低资源语种迁移, 可以采用多语种联合训练共享部分参数的方式来进行知识共享, 以对多个语种同时进行建模. 在目标领域模型自适应方面, 尽管神经网络模型自适应技术取得了巨大的进展, 但是在神经网络模型复杂、关于目标领域数据的先验知识(领域类别、语言学知识等)缺乏、自适应算法的鲁棒性等方面还有诸多问题. 如何在低资源的目标领域数据上对通用领域模型进行自适应, 提升目标领域语音识别性能具有挑战.

对于不同信道间知识迁移的研究方法可以采用不同信道的语音同时训练的策略, 需要研究如何使用丰富资源信道模型指导匮乏资源信道的模型训练.

(iii) 多源数据信息综合利用

在已有成对数据基础上, 挖掘音频、文本、视频等数据中的有效信息. 通过无监督学习、半监督学习以及自学习等学习策略的综合, 构建一套性能超越单系统并具有一定的机器持续自学习能力的新一代语音识别系统.

(2) 发展基于人类听觉机制的语音处理和识别技术, 实现语音前端处理、语音内容、语种、说话人和

性别的统一识别, 减少对开源算法的依赖.

结合类脑机制的研究, 探索人类听觉机制, 例如听觉特征提取机制、听觉注意力机制(包括自上而下的有意识的注意力和自下而上的无意识的注意力)等, 并将此机制应用于语音前端处理、语音识别建模、模型自适应等方面.

根据在多说话人场景中, 人类听觉系统会利用大脑中的长期记忆和对当前环境的感知信息实现对感兴趣目标说话人语音的关注这一特性, 提出在基于深度神经网络的目标语音提取算法中, 利用说话人特征编码模型, 从注册语音中获取长期记忆特征, 并与从初步提取的混合语音特征中获取的短期记忆特征进行融合.

此外, 传统的语音前端处理、语音识别、说话人识别、语种识别和性别识别是分别处理的, 未来可以发展统一的框架, 实现这些信息的统一识别. 上述模块整合到统一框架后, 需要针对不同任务, 对不同特征进行选择组合, 适配目标任务. 在任务协同训练时, 针对各个场景设计联合训练的目标函数.

(3) 优化识别算法, 减少对于算力的依赖.

优化识别算法, 采用知识蒸馏、模型剪枝、模型量化等技术, 采用端到端非自回归解码等技术, 通过异构协同解码, 实现显著加速效果.

(4) 实现在国产自主硬件设备上的智能声学技术.

采用高速缓存优化、超标量计算优化、异构计算优化等技术, 集成CPU加速算子库、GPU加速算子库. 实现在国产硬件设备上的全栈自主可控的声学智能化技术, 减少对于国外硬件CPU和资源的依赖.

(5) 提升语音系统的安全性.

为减少平台中的安全隐患, 技术人员需要对所有的第三方工具的源代码进行详细的了解. 训练平台需要保证在运行环境上做到自主可控, 同时保证代码的安全可靠. 在自主可控性上, 平台不依赖于任何外界商业化软件, 并尽可能的减少第三方工具的使用.

构建更加鲁棒的语音识别系统, 实现语音对抗样本检测, 探索基于集成学习的对抗检测方法, 利用集成学习技术, 增加系统的复杂度和模型结构的多样性. 结合差分隐私等技术加密保护模型输出, 探索数据信息反窃取算法. 通过Dropout等技术向模型输出引入噪声不确定性干扰, 从而难以进行反向推演, 实现具备隐私保护功能的模型防御机制, 抵抗特定的推断攻击.

(6) 将人工智能技术大规模引入医疗声学方向, 减少对开源算法的依赖.

在特征提取方面, 可利用GABOR滤波器替代传统的STFT, 以更好地在时间与频率局部化上进行折中; 利用可学习的低通池化模块代替传统的mel滤波器组, 使模型更为自由地选取潜在独特特征; 利用sPCEN特征归一化模块, 代替传统的固定的对数归一化模块, 使得模型更加贴近人类对于音量的非线性感知.

可以采用基于深度神经网络的自监督声学特征提取器, 利用大量无标签咳嗽声数据提取咳嗽声内在特征规律; 利用wav2vec卷积网络进行对比预测学习, 利用context网络预测encoder网络帧级输出, 进而获取语音内部内在规律.

由于医疗声学的训练数据有限, 分类类别少, 可以采用基于轻型卷积神经网络的小样本学习技术; 采用LightCNN架构, 结构简单, 避免过拟合; 采用提升鲁棒性的损失函数与正则化方法等.

参考文献

1 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444

2 Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks. Pattern Recogn, 2018, 77: 354–377

3 Lecun Y, Bengio Y. Convolutional networks for images, speech, and time series. The Handbook of Brain Theory and Neural Networks. Cambridge: MIT Press, 1998. 255–258

4 Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks, 2015, 61: 85–117

5 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput, 1997, 9: 1735–1780

6 Zhang S, Zheng D, Hu X, et al. Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Shanghai, 2015. 73–78

7 Pan S J, Yang Q. A survey on transfer learning. IEEE Trans Knowl Data Eng, 2009, 22: 1345–1359

8 Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning. arXiv: 1312.5602

9 Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, 2014

10 Young S, Evermann G, Gales M, et al. The HTK book. Cambridge Univ Eng Depart, 2002, 3: 12

11 Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit. In: Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. Hawaii, 2011

12 Watanabe S, Hori T, Karita S, et al. Espnet: End-to-end speech processing toolkit. arXiv: 1804.00015

13 Ying D, Yan Y, Dang J, et al. Voice activity detection based on an unsupervised learning framework. IEEE Trans Audio Speech Lang Process, 2011, 19: 2624–2633

14 Baevski A, Auli M, Mohamed A. Effectiveness of self-supervised pre-training for speech recognition. arXiv: 1911.03912

15 Baevski A, Zhou Y, Mohamed A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst, 2020, 33: 12449–12460

16 Liu A T, Li S-W, Lee H-Y. Tera: Self-supervised learning of transformer encoder representation for speech. IEEE/ACM Trans Audio Speech Lang Process, 2021, 29: 2351–2366

17 Baevski A, Schneider S, Auli M. vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv: 1910.05453

18 Hsu W-N, Bolte B, Tsai Y-H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans Audio Speech Lang Process, 2021, 29: 3451–3460

19 Chung Y-A, Zhang Y, Han W, et al. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. arXiv: 2108.06209

20 Lai C-I, Chuang Y-S, Lee H-Y, et al. Semi-supervised spoken language understanding via self-supervised speech and language model pretraining. In: Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, 2021. 7468–7472

21 Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv: 1810.04805

22 Zhang Y, Qin J, Park D S, et al. Pushing the limits of semi-supervised learning for automatic speech recognition. arXiv: 2010.10504

23 Zhang Y, Park D S, Han W, et al. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. arXiv: 2109.13226

24 Inaguma H, Cho J, Baskar M K, et al. Transfer learning of language-independent end-to-end ASR with language model fusion. In: Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, 2019. 6096–6100

25 Khare S, Mittal A, Diwan A, et al. Low resource ASR: The surprising effectiveness of high resource transliteration. Interspeech, 2021, 1529–1533

26 Kannan A, Datta A, Sainath T N, et al. Large-scale multilingual speech recognition with a streaming end-to-end model. arXiv: 1909.05330

27 Liu D, Wan X, Xu J, et al. Multilingual speech recognition training and adaptation with language-specific gate units. In: Proceedings of the 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP). Taipei, 2018. 86–90

28 Zhao Z, Elshamy S, Fingscheidt T. A perceptual weighting filter loss for DNN training in speech enhancement. In: Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). New Paltz, 2019. 229–233

29 Han W, Zhang X, Min G, et al. Perceptual weighting deep neural networks for single-channel speech enhancement. In: Proceedings of the 2016 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China. 2016. 446–450

30 Erdogan H, Hershey J R, Watanabe S, et al. Improved mvdr beamforming using single-channel mask prediction networks. Interspeech, 2016, 1-5: 1981–1985

31 Sun X, Xia R, Li J, et al. A deep learning based binaural speech enhancement approach with spatial cues preservation. In: Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, 2019. 5766–5770

32 Fan L. Audio example recognition and retrieval based on geometric incremental learning support vector machine system. IEEE Access, 2020, 8: 78630–78638

33 Chen H, Zhang P, Bai H, et al. Deep convolutional neural network with Scalogram for audio scene modeling. Interspeech, 2018, 1-6: 3304–3308

34 Alamir M A. A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers. Appl Acoust, 2021, 175: 107829

35 Serdyuk D, Audhkhasi K, Brakel P, et al. Invariant representations for noisy speech recognition. arXiv: 1612.01928

36 Park J-H, Oh M, Park H-M. Unsupervised speech domain adaptation based on disentangled representation learning for robust speech recognition. arXiv: 1904.06086

37 Higuchi Y, Watanabe S, Chen N, et al. Mask CTC: Non-autoregressive end-to-end ASR with CTC and mask predict. arXiv: 2005.08700

38 Higuchi Y, Inaguma H, Watanabe S, et al. Improved mask-CTC for non-autoregressive end-to-end ASR. In: Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, 2021. 8363–8367

39 Gao C, Cheng G, Zhou J, et al. Non-autoregressive deliberation-attention based end-to-end ASR. In: Proceedings of the 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP). Hong Kong, 2021. 1–5

40 Chorowski J, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition. arXiv: 1506.07503

41 Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, 2016. 4960–4964

42 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, 2017. 6000–6010

43 Miao H, Cheng G, Zhang P, et al. Online hybrid CTC/attention end-to-end automatic speech recognition architecture. IEEE/ACM Trans Audio Speech Lang Process, 2020, 28: 1452–1465

44 Moritz N, Hori T, Le J. Streaming automatic speech recognition with the transformer model. In: Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, 2020. 6074–6078

45 Miao H, Cheng G, Zhang P, et al. Online hybrid CTC/attention architecture for end-to-end speech recognition. Interspeech, 2019, 2623–2627

46 Sotelo J, Mehri S, Kumar K, et al. Char2wav: End-to-end speech synthesis. Proceedings of the 2017 5th International Conference on Learning Representations (ICLR). Toulon, 2017

47 Mehri S, Kumar K, Gulrajani I, et al. SampleRNN: An unconditional end-to-end neural audio generation model. arXiv: 1612.07837

48 Ping W, Peng K, Chen J. ClariNet: Parallel wave generation in end-to-end text-to-speech. arXiv: 1807.07281

49 Ren Y, Hu C, Tan X, et al. FastSpeech 2: Fast and high-quality end-to-end text to speech. arXiv: 2006.04558

50 Donahue J, Dieleman S, Binkowski M, et al. End-to-end adversarial text-to-speech. arXiv: 2006.03575

51 Weiss R J, Skerry-Ryan R J, Battenberg E, et al. Wave-tacotron: Spectrogram-free end-to-end text-to-speech synthesis. In: Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, 2021. 5679–5683

52 Hampshire J B, Waibel A H. A novel objective function for improved phoneme recognition using time-delay neural networks. IEEE Trans Neural Netw, 1990, 1: 216–228

53 Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks. IEEE Trans Acoust Speech Signal Process, 1989, 37: 328–339

54 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016. 770–778

55 Cheng G, Peddinti V, Povey D, et al. An exploration of dropout with LSTMs. Interspeech, 2017, 1-6: 1586–1590

56 Povey D, Cheng G, Wang Y, et al. Semi-orthogonal low-rank matrix factorization for deep neural networks. Interspeech, 2018, 1-6: 3743–3747

57 Cheng G, Povey D, Huang L, et al. Output-gate projected gated recurrent unit for speech recognition. Interspeech, 2018, 1-6: 1793–1797

58 Miao H, Cheng G, Gao C, et al. Transformer-based online CTC/attention end-to-end speech recognition architecture. In: Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, 2020. 6084–6088

59 Zhang S, Liu C, Jiang H, et al. Feedforward sequential memory networks: A new structure to learn long-term dependency. arXiv: 1512.08301

60 Zhang S, Lei M, Yan Z, et al. Deep-FSMN for large vocabulary continuous speech recognition. In: Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, 2018. 5869–5873

61 Tan K, Chen J, Wang D. Gated residual networks with dilated convolutions for monaural speech enhancement. IEEE/ACM Trans Audio Speech Lang Process, 2019, 27: 189–198

Thinking and exploration into intellectualization of speech and medical acoustics

YAN YongHong[1,2*] & CHENG GaoFeng[1,2]

[1] *Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China;*
[2] *University of Chinese Academy of Sciences, Beijing 100049, China*

As the artificial intelligence boom sweeps across all walks of life, researchers are also intellectualizing the area of acoustics. Taking speech and medical acoustics as examples, we first analyzed the significant challenges faced by acoustic intelligence, including data scarcity, algorithm independence, computing power dependence, security, and talent shortage. We also summarized important relevant technology and discussed prospective remedies. Furthermore, we reviewed the recent major scientific advances in speech and medical acoustics, both at home and abroad, and summarized the development history of the domestic intelligent voice industries. Finally, we look forward to the directions and trends of intellectual speech and medical acoustics and provide suggestions for specific issues in these fields.

speech acoustics, acoustic intelligence, medical acoustics

PACS: 43.60.Np, 43.71.Sy, 43.72.Bs, 43.72.Ne, 43.80.Qf

doi: 10.1360/SSPMA-2021-0306