特色专题

# 人工智能在声学中的应用及展望

郑成诗[1,2],李安冬[1,2],饶丹[3],袁旻忞[4],江峰[1],李晓东[1,2]

**摘要** 人工智能(artificial intelligence,AI)正与声学中的水声学、超声学和空气声学深度交叉融合,持续推动声学技术的革新。本文重点探讨AI在声学中的应用,尤其是其在空气声学领域中的应用。首先,详细阐述其在语音信号处理、声源定位、空间音频、环境声检测、分类与智能监测以及声学仿真与优化等方面的应用现状,并分析其相较于传统方法所具备的优势。在语音信号处理领域,AI已实现从特征工程到端到端建模的范式转变。其中,基于深度学习的语音识别、增强和合成技术,不仅在特定任务中超越了人类水平,还通过多模态融合和生成式模型拓展了应用边界。然后,针对应用过程中可能出现并导致其难以满足实际应用需求的核心问题展开讨论,包括泛化性、数据依赖与质量、复杂度、实时性及多模态融合问题。最后,总结了AI在声学应用中所面临的挑战和未来的发展方向。在基础理论层面,声学与AI的交叉研究尚未建立完善的理论框架,需要重点开展研究,以指导模型设计和性能评估。在技术层面,如何平衡算法复杂度与性能,实现可扩展的实时处理,仍是亟待解决的关键技术难题。未来,“AI+声学”将在海洋探测、医疗诊断、虚拟现实、环境声学等领域进一步发挥重要作用,最终实现从实验室研究、单点技术落地到大规模产业化应用的跨越。

**关键词** 人工智能;深度学习;声学;音频信号处理;声源定位;空间音频

声学[1−2]作为物理学的一个重要分支,按照声波传播媒介分为水声学、超声学和空气声学,分别研究声波在液体中(水中)、固体中和空气中的科学问题与实践应用[3]。水声学[4]涵盖物理水声学、水声学工程、水声通信与导航、水声探测与成像和海洋生物声学等多个分支学科,在海洋探测、通信导航和海洋生态保护方面有着重要的应用价值。超声学[5]包括基础超声学、医学超声学、工业与工程超声学、超声电子学与信号处理等多个分支学科,在精准医疗和智能检测等方面有着重要应用。空气声学[3]与人类的日常生活息息相关,包括音乐声学、语言声学、建筑声学、环境声学、电声学、心理与生理声学等多个分支学科,其中音乐声学在声学领域的研究中历史最为悠久。传统音乐声学[6]主要研究乐器发声原理、乐音传播特性与听觉感知规律等方面,近年来,三维空间声音合成与音乐自动生成则成为新的研究热点;传统语言声学主要研究语音产生机制、传播规律及感知特性等方面,语音增强、语音识别和语音合成则成为新的研究热点。

相较于声学的悠久历史,AI[7−8]则是始于20世纪40年代的新兴交叉学科,已与计算机科学、数学、神经科学等领域深度融合。早期AI旨在让机器模拟人类处理特定任务,因而被称为狭义人工智能(narrow

1. 中国科学院声学研究所,噪声与音频声学实验室,北京 100190
2. 中国科学院大学,北京 100049
3. 华南理工大学物理与光电学院,广州 510641
4. 交通运输部公路科学研究院,北京 100088

收稿日期:2025−06−23;修回日期:2025−11−22
基金项目:国家重点研发计划项目(2021YFB3201702)
作者简介:郑成诗,研究员,研究方向为通信声学,电子信箱:cszheng@mail.ioa.ac.cn
引用格式:郑成诗,李安冬,饶丹,等. 人工智能在声学中的应用及展望[J]. 科技导报,2026,44(4):62−78;doi:10.3981/j.issn.1000-7857.2025.09.00081

artificial intelligence, NAI)或弱人工智能(weak artificial intelligence, WAI)。近年来,以深度学习(deep learning, DL)[9]为核心技术路线的NAI已经在视觉和听觉等领域接连取得突破,甚至部分任务(如中英文语音识别[10−11])性能已超越人类水平。随着DL的飞速发展,能够在众多领域模拟人类处理问题的通用人工智能(artificial general intelligence, AGI)应运而生,成为科技巨头竞争的焦点,同时也已成为大国综合国力竞争的制高点。值得一提的是,AGI已初步展现出处理复杂问题的强大能力,但其模型参数量达数百亿至万亿级,现阶段通常运行于云服务器;随着芯片技术的进步,单芯片处理能力不断突破,AGI也开始被部署于单机系统,实现本地化应用。介于NAI和AGI之间的行业AI大模型也层出不穷,涉及医疗、传媒、教育、金融等行业,相比于AGI,在特定行业更具专业性和经济性。尽管3者在模型复杂度、专业性以及经济性方面存在着较大差异,但其技术内核,如所采用的网络架构等方面并无显著性差异,因而下文不区分3者,统称为AI。

AI与声学的结合始于20世纪50年代,早期主要应用于语音识别和语音合成领域,典型代表为贝尔实验室在1952年开发的首个人工语音识别系统“Audrey”,通过提取共振峰频率作为声学特征并结合模式匹配算法实现了英文数字0~9的识别。经过70余年的融合发展,其在声学领域的应用已超越语音信号处理范畴,深度融入语音处理、声源定位、空间音频、声学场景检测与分类及声学仿真与优化等多个分支领域,全面推动水声学、超声学和空气声学的技术革新,显著提升各分支领域的性能表现。

本文聚焦AI在声学中的应用,重点阐述“AI+声学”技术的发展现状,并与传统声学技术展开对比分析,剖析该领域所面临的多重挑战,并对“AI+声学”技术的未来发展方向进行展望。

## 1 基本概念

### 1.1 声学基本概念

声学主要研究声音的产生、传播、接收和效应等。一类重要的声音即自然语音特指人类发声器官发出的,具有特定意义的声音,是人类交流的核心工具。不少动物也有自己的声音语言,但是由于与人类发声器官存在差异,与人类发出的自然语音具有显著差异。噪声,如工业噪声、建筑施工噪声、交通运输噪声及社会生活噪声则分别是由工业生产活动中使用固定设备产生的噪声、施工过程机械作业产生的噪声、各类机动车辆和船舶等交通工具运行时产生的噪声及其他人为活动所产生的噪声。不同类型的声音通常在频率、声压级、频谱等物理属性上存在差异,在心理属性上则对应于音调、响度和音色的差异。考虑声音产生后在空间中进行传播,不同声音通常还存在空间位置差异,对应于声音在心理属性上还存在方向感和环绕感差异。

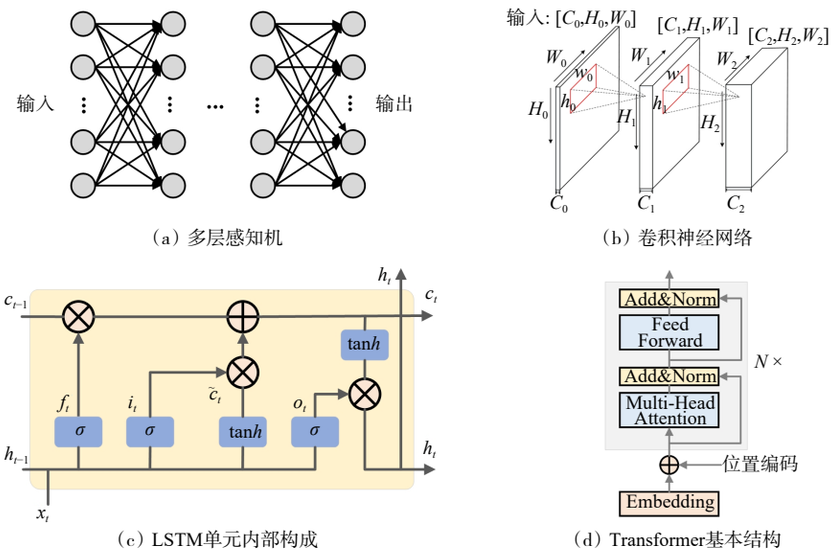
对在空间中传播的声音进行接收、处理、识别和定位及调控是当前声学的研究热点。对语音进行捡拾、定位、增强、识别和合成,已成为声学领域最为重要的研究分支之一;对其他类型的声音信号进行有效捡拾、处理、检测和定位,也得到了广泛的关注和研究。近年来,通过设计声学超结构超材料实现声波主动调控成为新的研究热点。

### 1.2 AI基本技术

AI技术源于对人类智能的模拟与探索,其早期演进为现代核心模型架构奠定了重要基础。1943年,McCulloch等[12]提出首个神经元数学模型,构成人工神经网络的理论雏形;1958年,感知机模型的出现开启了神经网络的热潮[13],虽因线性局限陷入低谷,但为后续发展累积了关键经验。1986年,误差反向传播(back propagation, BP)算法的提出,解决了多层神经网络的训练难题[14],推动多层感知机(multi−layer perception, MLP)成为早期核心模型;2006年,Bengio等[15]提出的分层预训练方法,有效突破了深度网络的优化瓶颈,标志着AI技术正式复兴,为后续复杂模型研发提供了核心思路。

经过数十年演进,AI领域形成了以经典模型为核心的技术体系,各模型凭借独特结构适配不同数据处理需求,如图1所示。MLP作为最基础的全连接神经网络,由输入层、隐藏层、输出层构成,通过神经元权重密集连接及激活函数形成非线性映射,可学习数据全局统计特征,适用于维度较低、特征分布规整的分类与回归任务,是复杂模型的基础架构支撑。

卷积神经网络(convolutional neural network, CNN)




（a）多层感知机　（b）卷积神经网络

（c）LSTM单元内部构成　（d）Transformer基本结构

图 1　不同 AI 基本模型结构示意

以“局部感知+参数共享”为核心，通过卷积核捕获数据局部关联特征，在减少参数量的同时保留关键信息。其结构在局部相关性数据处理中表现突出，可高效提取层级化特征，适配多通道信号、图像等数据类型。从早期 LeNet−5[16]到 AlexNet[17]和 ResNet[18]等，该类结构的演化持续提升其特征提取能力。

长短期记忆网络(long short−term memory, LSTM)[19]是适配时序数据的循环神经网络(recurrent neural network, RNN)变体，通过遗忘门、输入门、输出门的门控机制，解决传统 RNN 面临的长序列梯度消失(gradient vanishing)问题。其可选择性记忆历史有效信息、更新信息并调节输出权重，能有效捕获长时程时序数据动态关联；变种门控记忆单元(gated recurrent unit, GRU)由 Cho 等[20]提出，相比 LSTM 进一步简化了结构，提升了运算效率。

Transformer 模型于 2017 年由 Google 提出[21]，核心为自注意力(self−attention, SA)机制，通过计算数据不同位置间的关联权重，实现全局信息的并行捕获与动态聚焦。相较于 LSTM 的时序渐进式处理，其可同步整合全局上下文信息，强化关键信息的特征表达，在长时长、高冗余的复杂数据处理中优势显著。该模型也已逐渐成为当前大语言模型(large language models, LLM)与扩散模型(diffusion models, DM)的标准结构。

这些核心技术从早期简单模型逐步演进为复杂高效的架构体系，为不同类型数据的解析提供了针对性工具，与声学概念互补，共同支撑起了“AI+声学”的融合应用与创新发展。

## 2　应用现状

从技术应用现状来看，传统声学处理整体围绕 5 大核心需求展开，为后续细分应用场景提供基础支撑：针对语音信号的“采集−特征提取−语义解析”需求，奠定语音信号处理中语音增强、识别、合成等的技术基础；针对声音空间位置确定的需求，形成声源定位技术的核心目标，通过传播时间差和相位差等特征，实现多场景下声源的精准定位；针对声场空间信息还原与个性化体验需求，围绕声场重构、音频上混、头相关传递函数(head−related transfer function, HRTF)个性化等空间音频技术展开；针对环境中有效噪声信号与分类的需求，形成噪声智能检测与分类技术的核心目标，构建环境声分类与检测等技术基础；针对声学组件性能评估与结构设计需求，围绕物理参数建模与仿真计算，实现声学结构的高效设计与性能优化。

### 2.1　AI+语音信号处理

语音信号处理作为 AI 与声学交叉融合的核心领域，其发展历程见证了从统计信号模型到数据驱动范式的转变[22]。当前，学术界与工业界聚焦于深度模型表征能力提升、多场景鲁棒性优化及跨模态信息融合，相关成果不断推动语音处理技术在基础理论与工

程应用上的双重突破。

早期的语音信号处理技术如语音识别主要基于隐马尔可夫模型(hidden Markov model, HMM)处理常用语音特征, 如梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)。因模型建模与特征表征能力有限, 早期语音识别技术在复杂声学环境中泛化能力不足。随着 AI 技术的兴起, CNN 凭借分层时频特征提取能力展现出显著优势, 而 LSTM 因自回归处理模式, 契合语音分帧处理模式和时间序列信号特性, 被广泛应用于各类语音信号处理任务中。目前, 主流语音信号处理任务主要涵盖语音识别、语音增强和语音合成。

### 2.1.1 语音识别

在语音识别领域, 2012 年, 微软研究院与 Google 率先使用深度神经网络(deep neural network, DNN), 将语音识别错误率降低 20%~30%。这一突破标志着语音识别从依赖手工设计特征与传统模型, 转向基于数据驱动的 DNN 范式。此后, 学界与业界围绕 DNN 优化展开深入研究。2015 年, 百度公司提出 Deep-Speech 2[10], 在中文语音识别任务中首次超越人类专业速记员水平。随着 Transformer 架构在自然语言处理领域的推广, 基于 Transformer 的预训练与微调范式也由此逐步在语音识别领域广泛应用: 通过自监督学习在海量无标注语音数据上进行预训练, 学习语音信号的通用特征表示, 再利用少量标注数据微调适配下游任务, 这种模式在低资源语种识别中显著降低词错误率, 极大缓解了对大规模标注数据的依赖。2023 年, OpenAI 推出了 Whisper 模型[23], 通过大规模预训练构建多语种通用识别能力, 再经高效微调适配特定场景, 实现了语音识别、转写与翻译的一体化突破, 显著提升了复杂环境与低资源语种识别的鲁棒性。

随着研究深入, 多模态融合的语音识别技术成为新的研究热点。在语音与视觉联合建模中, 研究人员通过摄像头捕捉说话人的唇动信息, 利用跨模态特征融合技术实现唇动特征与语音特征的精准对齐(图 2)[24]。相较于单模态场景, 这种多模态融合技术在强噪声和强混响声学环境中能有效降低词错率, 显著提升语音识别的鲁棒性。近年来, 脑机接口与语音识别的融合也取得了突破性进展, 例如, Kamble 等[25]尝试结合脑电图(electroencephalogram, EEG)信号进行语音识别,

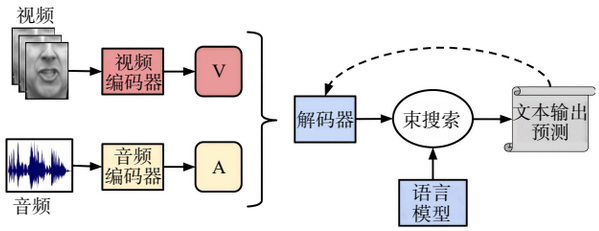



图 2 唇动-语音多模态识别系统示意

取得了一些进展。但由于脑电信号具有高噪声、个体差异大等复杂性, 该技术在性能与泛化性方面仍存在巨大的研究提升空间。

### 2.1.2 语音增强

在语音增强领域, 传统基于统计信号处理的方法如谱减法、维纳滤波等, 受限于对噪声先验的依赖, 在非平稳噪声、强混响等复杂场景下性能显著下降。随着深度学习技术的发展, 数据驱动的语音增强技术通过构建带噪与纯净语音的非线性映射关系, 实现了从“模型假设”到“数据学习”的范式转变[22, 26]。2014 年, Xu 等[27]利用多层 DNN 学习带噪语音对数谱到干净对数谱的非线性映射(图 3), 相比传统算法, 其在各项指标上实现显著提升。后续研究结合 CNN 与 LSTM 的处理特点与优势, 在网络架构上持续探索, 进一步提升了性能[28]。近年来, 一系列相关比赛也为该方向的发展注入新的活力, 微软公司于 2020 年发起深度降噪(deep noise suppression, DNS)挑战赛, 迄今已连续举办 5 年, 该比赛的举办进一步推动了该领域进步。随着生成式与大模型技术的蓬勃发展, 基于 LLM 和 DM 的语音增强方法成为新的研究热点[29]。此外,

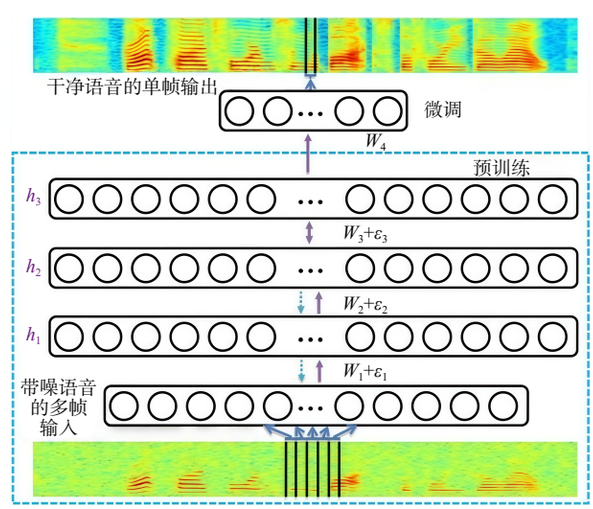



图 3 基于 DNN 的语音增强示意

跨模态的语音增强技术通过融合视频与音频线索,在相对恶劣的声学场景下取得较单一音频模态更好的性能[30]。然而,面对实际场景中可能出现的模态缺失问题,如何提升多模态语音增强方法的鲁棒性,仍是亟待解决的关键问题。

### 2.1.3 语音合成

语音合成技术经历了从参数化统计语音合成到端到端生成的跨越式演进。早期参数合成方法,如STRAIGHT[31]等,通过提取基频、谱包络等核心声学参数,结合规则驱动的波形生成机制实现语音合成,虽能满足基础语音输出需求,但普遍存在韵律僵硬、自然度欠佳等问题。随着 AI 技术在生成式建模领域的突破,语音合成实现了从“参数驱动”到“数据驱动”的根本性转变。2016 年,DeepMind 推出的 WaveNet 模型[32]取得了合成语音质量的突破性进展(图 4),其通过因果扩张卷积对语音波形概率分布进行深度建模,显著提升了合成语音的自然度,平均意见得分(mean opinion score, MOS)从传统参数化合成方法的 3.6 提升至 4.0 以上。但 WaveNet 的串行计算特性导致合成效率低下,难以适配实时应用场景。Transformer 架构的提出进一步革新了语音合成技术范式。浙江大学研究人员提出 FastSpeech 系列工作[33],通过时长预测与声学特征解耦的设计,在保持高自然度的同时,将语音合成速度提升了数十倍。近年来,LLM 和 DM 为语音合成带来了新的技术突破。基于 LLM 的方法借助海量文本−语音数据与强大的 Token 级预测能力,在零样本(zero−shot)语音合成任务中取得显著提升[34];DM 则通过逐步去噪的反向生成机制,逐步生成多样化与高质量的音频。此外,多模态融合与个性化合成已成为当前研究热点,为定制化语音生成与高表现力语音提供了可能,在娱乐、人机交互等领域具有重要应用价值。

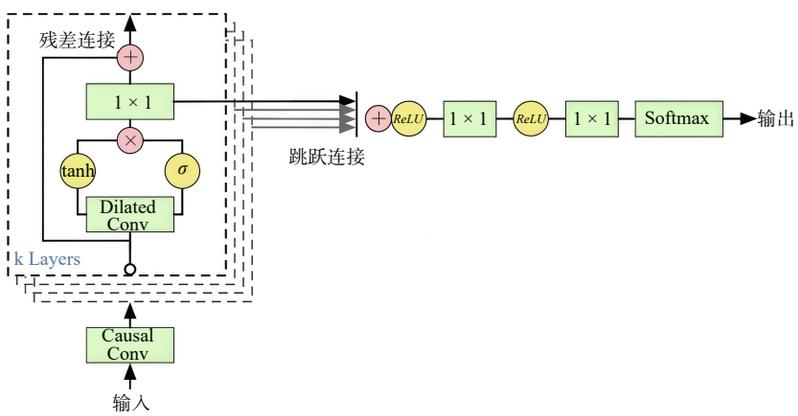



图 4 WaveNet 结构示意

## 2.2 AI+声源定位

人类仅用单耳就可实现声源定位,源于不同方向入射的声源受头部、躯干及耳廓等的散射与滤波效应差异[35];受此仿生启发,有研究人员将单个传声器嵌入预设计的三维超材料结构中,实现了多声源实时定位和分离[36]。研究表明,由于不同入射方向的声源到达双耳存在时间差(interaural time difference, ITD)和声级差(interaural level difference, ILD),因而相较于单耳,双耳可更准确地定位声源[37]。现有的绝大多数声源定位系统均通过多传感器拾取信号,并利用时延等特征估计声源位置。

1916 年,法国科学家 Paul Langevin 发明了首台实用化声呐定位系统,可同步测定目标方位与距离。水下声呐通常分为被动式和主动式,前者通过被动接收目标辐射噪声的特点来确定目标位置,后者则主动发射声音并接收回声来探测和定位目标[38]。1794 年,意大利科学家 Lazzaro Spallanzani 研究了蝙蝠进行空间定位的基本机制,证实其不依赖视觉导航。1913 年,Richardson[39]基于超声波原理发明了回声定位器,奠定了主动式超声定位基础[40]。20 世纪 10 年代,空气声学定位方法兴起,早期基本都是通过设计声学结构实现集音器的效果,并用于单人或多人的双耳来判断和定位飞艇或飞机等目标。尽管早年的声源定位系统大多源于军事用途,但如今已广泛应用于海洋通

信导航、医学诊断、消费电子等民用领域。

#### 2.2.1 传统声源定位方法

传统声源定位方法包括可控波束响应(steered response power of a beamformer, SRP)[41]、基于高分辨率谱估计[42]和基于时间差(time difference of arrival, TDOA)[43]等方法。

可控波束响应的典型方法为延迟相加波束形成(delay-and-sum beamforming, DSB),这类方法通常需要预先计算某一方向声源的每个频带两两传感器之间的传播时延,补偿传播时延后求和所有频带所有两两传感器之间的互功率谱;再搜寻全空间所有方向的最大值以定位声源方位。在噪声和混响环境下,可控波束响应方法性能显著下降,需针对性加权优化:纯噪声环境中,估计每个频带的信噪比,并以此为依据加权[43];纯混响环境中,估计每个频带的相位变换(phase transform, PHAT),并以此加权[44]。由于可控波束相应方法需要在全空间进行波束扫描,并搜寻最大值,因此该类方法运算复杂度较高。

基于高分辨率谱估计的定位方法,包括最小方差(minimum variance, MV)谱估计和基于特征值分析(eigenanalysis-based)的方法如 MUSIC(multiple signal classification)、ESPRIT(estimation of signal parameters via rational invariance techniques)以及 MODE(method of direction estimation)等算法[45]。这类方法通常需要首先估计空间相关矩阵,且假定声源具备统计平稳,当声源位置移动或者声源二阶统计特性不平稳如语音信号,这类算法的定位性能会呈现不同程度的退化。尽管基于高分辨率谱估计的定位方法相比于常规波束形成如 DSB 具有更高分辨率,在理想情况下也具有更高的定位性能,但是对信号建模误差如传声器位置误差和传感器一致性误差等鲁棒性不足。早期的高分辨率谱估计定位方法仅适用于窄带信号,但这类方法已被扩展到宽带信号,也可应用于多声源定位场景。相比于可控波束响应方法,基于高分辨率谱估计的定位方法每次迭代所需要的运算复杂度更低。

基于 TDOA 的定位方法有 2 个阶段:第一阶段估计任意 2 个传感器接收信号的相对时延,第二阶段根据传感器的相对位置以及第一阶段估计得到的相对时延通过解一组非线性方程得到声源位置的极大似然估计值。这类方法的性能取决于第一阶段估计的相对时延的准确性,Knapp 等[43]提出的广义互相关(generalized cross-correlation, GCC)方法是应用最为广泛的相对时延估计方法。GCC 与 SRP 方法一样容易受到噪声和混响等因素的干扰,同样可通过对互功率谱加权以提高相对时延估计性能。基于 TDOA 的定位方法有一个前提假设是单一声源,对于多声源场景需要用多段短时信号进行联合估计。已有研究结果表明,基于时间差的定位方法在多声源、强噪声或者中等混响以上声学场景性能不佳[46]。

#### 2.2.2 AI 声源定位方法

Grumiaux 等[46]对基于深度学习的室内声源定位进行了全面的总结,涵盖神经网络架构、输入特征与输出目标、训练及测试数据生成与获取途径,以及深度学习方法。

如图 5[46]所示,基于 AI 的声源定位方法的处理流程与传统 TDOA 定位方法类似,可分为 2 个阶段:第一阶段提取定位所需特征,第二阶段通过预训练模型映射输出声源位置。Krause 等[47]对比了不同输入特征的声事件检测与定位性能,包括基于幅度的通道间声级差、基于相位的通道间相位差,以及基于协方差矩阵或主特征向量的特征;此外,传统声源定位方法中的诸多中间物理量(例如 MV 谱、GCC-PHAT 等)也可作为输入特征。第二阶段通常采用主流的 MLP、CNN、Transformer 等网络结构或其组合形式以实现更高的定位精度[48]。

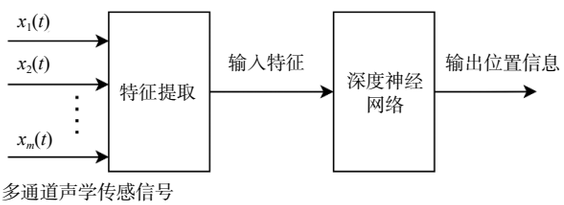



图 5 AI 声源定位处理流程

AI 声源定位方法的位置信息输出常采用分类或回归的方式:前者需将整个位置区域划分为多个子区域,通过神经网络输出各子区域的声源存在概率;后者则通过神经网络直接输出声源坐标信息。相较于分类方法给出的定位是离散值,回归方法给出的定位是连续值,因此其定位精度更高。然而,回归方法需预先知晓声源个数,且易出现声源置换问题。研究表

明,借鉴语音与音频分离中的置换不变性训练(permutation invariant training, PIT)策略,可有效解决回归方法中的声源置换问题[49]。由于在基于 AI 的声源定位中,输入特征与输出位置信息维度通常较低,神经网络架构相对简单,因而运算复杂度通常较低。

AI 声源定位方法通常需要大量训练数据优化模型参数,常用数据获取方式包括真实实验录制、仿真生成与数据增广 3 类。由于声源定位与声学传感器阵列的阵型高度相关,在室外应用中需考虑各类噪声影响,在室内应用则需兼顾噪声与不同混响时间的影响,因此录制真实场景典型实验数据的工作量极大。通过仿真生成训练数据是一种比较低成本的方案,其通常以声波传播模型为基础,结合射线声学、波动声学或 2 者融合的方式来生成声学传递函数,进而合成特定阵型传感器阵列的接收信号。研究表明,仅依赖仿真数据训练的模型在真实声学场景下存在鲁棒性不足的问题,而完全采用真实实验数据成本高昂,因此通过数据增广扩充有限真实数据成为折衷方案[50]。数据增广的实现方式因场景而异,简单方法包括加噪、滤波、通道置换等处理,复杂的方式则通过生成对抗网络(generative adversarial network, GAN)生成与真实数据难以区分的训练样本[51]。借助数据增广通常可提升 AI 声源定位系统在实际场景中的应用性能。

在学习方法方面,当前 AI 声源定位模型训练以有监督学习(supervised learning)为主,这种方法需依赖带标签训练数据[52];当训练数据缺少标签时,可采用半监督学习或弱监督学习策略。目前,基于 AI 的声源定位已应用于水下目标定位、空中目标定位及超声病灶定位等领域,在诸多场景中展现出优于传统方法的性能,因而具备潜在的研究价值与工程应用价值。

### 2.3 AI+空间音频

空间音频(spatial audio)旨在通过电声和信号处理手段,实现声场空间信息的捡拾、处理和重放,为听者提供身临其境的沉浸式听觉体验。作为声学、听觉心理和信息处理的交叉领域,空间音频技术已广泛应用于科学研究、消费电子、虚拟/增强现实等场景。近 20 年来,空间音频技术虽持续发展,但新兴应用场景与更高体验需求对传统技术提出新挑战。得益于 AI 在多领域的突破性成果,其在空间音频领域的应用探索也逐步推进,为解决传统技术瓶颈,提升性能及拓宽复杂场景适配范围提供了新路径。Cobos 等[53]讨论了 AI 技术在空间音频领域多个任务中的应用现状。

空间音频核心流程包括信号捡拾(合成)、处理和重放 3 个主要环节[54]:信号捡拾(合成)是指利用传声器阵列采集或根据虚拟声学场景合成包含声场空间信息的声音信号;处理环节是将捡拾(合成)的声音信号所包含的空间信息进行提取和修改的过程;重放环节则是通过扬声器或耳机重现处理得到的空间音频信号。目前,AI 技术主要应用于空间音频的信号处理环节,下面阐述相关的应用现状,重点放在取得了一定进展的方向,简要技术流程图如图 6 所示。

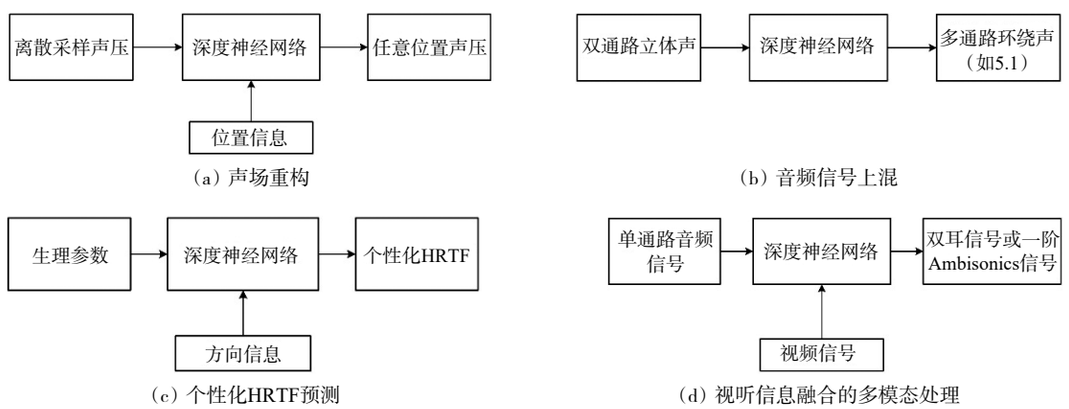



图 6 空间音频不同任务的 AI 应用技术流程示意

#### 2.3.1 声场重构

传声器采集的声场信号是空间离散的,而许多应用场合如声场空间信息分析、6 自由度声重放等需重构任意连续位置的声场信号。传统的声场重构方法

依赖线性内插或外插,其重构性能在采样密度较低时会显著下降。深度学习的引入为该问题提供了高效解决方案:受 CNN 在图像处理中的应用启发,Lluís 等[55]采用 U-Net 神经网络实现少量测量数据下矩形房间低频声场的重构。为使神经网络输出结果符合声场的物理特性,一些研究工作结合物理信息网络(physics-informed neural network, PINN)和声学信息网络(acoustic-informed neural network, AINN),融合波动方程作为物理约束[56]。得益于神经网络对声场共性统计特性强大的学习能力,在稀疏采样下,AI 方法的重构精度已逐步超越传统方法。

#### 2.3.2 音频信号上混

空间音频信号上混(upmixing)是将通路数较少的音频信号(通常为双通路立体声)转换为通路数更多的音频格式(如 5.1 通路环绕声)的过程,其技术本质在于拓展原音频信号的空间信息,例如在从 2~5 通路的上混处理中,增加了后方的环绕信息。传统的上混算法基于信号处理与心理声学原理,通过分析并分离双通路信号中的方向性成分和环境氛围成分,再经过一定的后处理后重新分配给上混后的各个通路。鉴于 AI 在声源识别与分离任务中的优势,相关技术被逐步应用于上混领域:Park 等[57]采用 DNN,以信号子带对数谱为输入,训练中央与环绕通路模型,实现立体声到 5.1 通路的转换;Choi 等[58]设计双 DNN 架构,分别负责信号分离与渲染,并将包含空间信息的通路 ILD 特征融入损失函数,强化网络空间信息提取能力。实验结果表明,AI 方法在主客观评价中均表现出较传统方法更优的性能,且该优势可能源于更强的方向性与环境氛围分离能力。

#### 2.3.3 个性化头相关函数预测

HRTF 是空间音频耳机重放的核心数据,具有显著个体差异性。个性化 HRTF 的测量需特定设备与场地,过程耗时费力。鉴于 HRTF 与人体生理参数(人头尺寸、耳廓外形等)高度相关,基于生理参数的个性化 HRTF 预测成为简便方案。传统的预测方法通常通过主成分分析等降维技术,分别降低 HRTF 以及生理参数的维度,然后建立降维后 HRTF 和生理参数之间的线性映射模型。利用此映射模型,就可以通过测量新个体的生理参数来预测其个性化 HRTF。为提升性能,相关研究利用 AI 技术来实现基于生理参数的个性化 HRTF 预测。Lee 等[59]提出一种 MLP-CNN 混合模型,利用 CNN 从耳部图像中提取耳廓生理参数,然后利用 MLP 预测个性化的头相关脉冲相应(head-related impulse response, HRIR)。Yao 等[60]提出一种基于变分自动编码器(variational autoencoder, VAE)的方法,并用于实现利用生理参数对 HRTF 的个性化预测。上述研究结果表明,借助 DNN 的非线性建模能力,基于 AI 的 HRTF 方法普遍取得了较传统线性映射更优异的性能。

#### 2.3.4 视听信息融合的多模态处理

传统空间音频技术通常仅利用音频信号进行单模态处理。一旦采集的音频信号本身空间信息缺失(如单个全指向传声器采集的信号),这种单模态处理方式往往难以恢复声源方向等关键空间信息。多模态 AI 技术通过融合视频信息补充缺失的空间维度,实现单通路音频到空间音频的生成。Gao 等[61]采用 U-Net 网络,从视频及对应单通路音频中生成双耳音频信号;Morgado 等[62]则利用 360°全景视频与自监督学习,通过生成时频掩码分离单通路信号中的方向性分量,进而将其编码为一阶 Ambisonics 信号。

迄今为止,AI 已在空间音频的多个任务中得到应用,并在特定场景下展现出优于传统方法的性能,凸显了其在空间音频领域的应用潜力。值得指出的是,AI 技术尚未在空间音频的所有任务中全面超越传统方法,主要原因在于空间音频涉及复杂的人类空间听觉机理,而当前对复杂场景下空间听觉的认知仍不完备;同时,高质量标准化训练数据的匮乏也是重要因素之一[53]。

### 2.4 AI+声学环境声检测、分类与噪声智能监测

声学环境声检测分类与监测以声学信号为核心研究对象,涵盖声学场景识别(如机场、教室、街道等)、特定声事件检测(如警报、玻璃碎裂、车辆行驶等)及环境噪声智能分析(如交通噪声、工业噪声、施工噪声等)等方向。旨在通过技术手段实现对复杂声学环境的精准感知与解读。作为声学、心理听觉与 AI 的交叉融合产物,该技术已广泛应用于智能安防、生态环境治理、智能家居、城市精细化管理等多个领域,成为支撑多场景智能化升级的关键基础技术[63-66],尤其在环境噪声污染防治等实际场景中发挥着不可替代的作用。

#### 2.4.1 传统机器学习方法

早期声学环境声检测、分类及监测工作,依赖人工设计声学特征(时域特征如过零率、短时能量,频域特征如频带能量,倒谱特征如 MFCC、梅尔频谱,及基于频谱图的 Gabor 滤波器组响应等)与浅层机器学习分类器,例如支持向量机(support vector machine, SVM)、HMM 等。这类方法虽在简单场景中取得一定效果,但面对复杂混合声学环境(例如城市中多源叠加噪声、相似声事件干扰)时,存在特征适应性差、抗干扰能力弱、泛化性能不足等局限,难以满足精准化、智能化的应用诉求。

#### 2.4.2 深度学习方法

AI 技术为声学环境声检测与分类带来了系统性革新,推动声学环境声检测、分类与噪声智能检测从“人工驱动”向“数据驱动”转型,基于 DL 的端到端技术路径逐渐成为主流。其核心优势体现在 3 方面:一是特征提取的自动化,模型通过 CNN、Transformer 等主流神经网络结构,以数据驱动方式自动挖掘声学信号中与监测和分类相关的关键特征,精准捕获不同场景、事件与噪声信号的细微时频差异;二是复杂环境的适配性,非线性表征能力让算法有效应对混合环境声、低信噪比等复杂场景,大幅度提升精度与分类鲁棒性;三是推动监测模式的智能化升级,实现噪声源实时定位、等效声级动态计算,改变传统监测依赖人工分析、效率低下的局面。

2016 年,IEEE SPS 发起的首次国际声学场景和事件检测及分类挑战赛(detection and classification of acoustic scenes and events, DCASE)挑战赛[67],成为该领域标准化与规模化发展的重要里程碑,推动了声学环境声检测与分类及噪声监测技术的快速迭代。早期,研究人员将 CNN 与 LSTM 结合,利用其局部特征提取与时序建模能力,在 DCASE 2017 任务中实现了 80% 以上的多场景分类准确率[68]。后续研究中,Gong 等[69]以时频谱为输入,结合 Transformer 建模长时特征关系,进一步提升了性能。

环境声的多样性、复杂性,以及高质量标注数据集的稀缺性,导致模型泛化能力面临挑战。预训练与迁移学习技术的应用有效缓解了这一问题。2017 年,Google 公司推出 AudioSet 数据集[70],其包含超 200 万条标注音频,覆盖 632 种音频事件,为模型的预训练提供了海量数据支撑。基于此,Kong 等[71]提出预训练音频模型 PANN,如图 7[71]所示,其基于 AudioSet 预训练,可灵活迁移至其他 6 种音频任务,并在声事件分类上取得当时最好的性能。

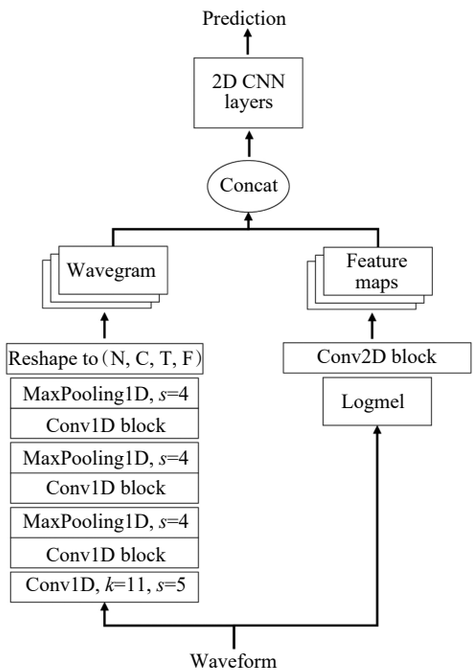



图 7 PANN 采用的预训练模型结构示意

此外,针对实际场景中噪声、小众声事件等标注稀缺问题,无监督、半监督及自监督学习方法陆续涌现,这类方法通过大量无标签声学数据中的潜在规律,有效扩展了技术在数据缺乏场景中的应用边界。在环境噪声监测领域,可用于未标注区域的噪声源聚类分析、设备故障噪声等突发噪声事件的自动检测;在工业场景中,可通过识别设备运行过程中的异常声信号(例如轴承磨损、管道泄漏等),实现故障早期诊断与决策。目前主流方案是结合大规模数据预训练与微调。其中,Han 等[72]基于 Transformer 架构进行预训练,然后通过低秩适配器(low-rank adaptation, LoRA)在工业异常检测数据集上完成微调,取得了比以往方案更好的性能。然而,由于通用声学事件数据与工业数据在特征模式上存在较大差异,该方向仍存在较大的进步空间。

### 2.5 AI+声学仿真优化

声学仿真与结构优化是声学工程中的关键任务之一,广泛应用于建筑声学、交通降噪、听觉增强设

备以及新型声学材料的研发。传统设计流程通常包括基于物理模型的正向建模(forward modeling)和参数扫描(parametric sweeping),其核心在于借助有限元法(finite element method, FEM)、边界元法(boundary element method, BEM)或传递矩阵法(transfer matrix method, TMM)等数值工具进行大量迭代仿真,以寻找满足目标性能的结构方案[73]。然而,随着问题复杂度的增加,如结构多样性提升、优化目标增多、设计空间高维化,传统方法逐渐暴露出不足:首先是计算成本高,难以实现实时反馈;其次是需大量专家经验指导,设计依赖性强;再次是缺乏目标性能到结构参数的高效反向路径;最后是对非凸、非线性以及多模态问题处理能力不足。因此,迫切需要一种更加灵活、高效、智能的优化设计范式。AI 技术的兴起为这一需求提供了一种可行路径,尤其在数据量不断积累、计算资源持续提升的背景下,AI 正逐步成为声学结构优化的重要引擎。

声学仿真通常涉及从结构参数到性能指标的映射,即正向问题(forward problem),而实际工程往往需要解决逆向问题(inverse problem),即从目标性能出发反推结构设计。这类问题高度非线性、维度高、解的空间复杂,传统的遗传算法(genetic algorithm, GA)、粒子群算法(particle swarm optimization, PSO)等虽然可用于全局优化,但计算开销大且收敛不稳定。随着 DL 和强化学习(reinforcement learning, RL)的兴起,研究者开始借助 AI 技术在复杂系统中实现反向设计、高维参数映射以及快速性能预测。AI 技术的引入为该领域带来了范式变革,基于 DL 的模型能够在目标性能到结构设计的反向路径中构建近似映射,实现更高效的结构生成与优化。

#### 2.5.1 传统声学仿真与优化方法

有限元法是声学结构分析中应用最广泛的数值技术,能够求解复杂边界条件下的声场分布。然而其缺点也非常突出:每次更改结构参数都需重新建模和求解,导致大量冗余计算。尤其在高维参数空间中,传统参数扫描方法的“穷举式”搜索极其低效。例如,在带隙优化任务中,研究人员需构造数千甚至上万种周期单元结构,通过逐一仿真获得其色散关系,再通过筛选选择性能最优解。这一流程不仅耗时长,而且易错失全局最优解。此外,拓扑优化是一种数学驱动的结构形貌优化方法,常用于最大化某一目标(如吸收系数、阻抗匹配等)。典型方法如固体各向同性材料惩罚模型(solid isotropic material with penalization, SIMP)和 Level-set 等,在早期超材料研究中有广泛应用[74]。然而,这类方法往往依赖梯度信息,难以适应非线性材料行为,且不易扩展至多目标情形。为提升优化效率,部分研究结合了遗传算法、粒子群算法等启发式搜索方法,缓解了参数空间维度高带来的问题。但这些算法本质仍是黑盒搜索,计算效率低、收敛性差,且在复杂结构多目标优化中仍需大量仿真样本支持。

#### 2.5.2 AI 声学仿真优化

AI 反向优化的基本思想是利用神经网络拟合性能与结构参数之间的映射关系。Donda 等[75]指出 MLP 适用于低维连续参数预测,而 CNN 适合处理网格结构的拓扑优化问题,在预测声学带隙、吸声频率等方面效果显著。由于一个目标性能可能对应多组结构参数,传统判别模型难以建模这种一对多关系。条件 GAN 和 VAE 在结构生成中表现出色,能够在满足性能约束的前提下生成多样化设计方案。此外,RL 将结构参数作为智能体状态,优化其行为策略以最大化目标性能,也被应用于带隙最大化与自适应材料设计[76]。近年来兴起的 PINN[77]备受关注。该类模型将物理定律(如声波传播方程)嵌入损失函数中,使训练过程兼具数据驱动与物理约束,提高了泛化能力与物理一致性。例如,Schmid 等[78]将 PINN 应用于声学边界导纳估计,通过嵌入亥姆霍兹方程物理约束,仅依托含噪声压数据,在无需显式指定边界条件且不依赖正向模型情况下,即可精准学习声场分布并隐式反演边界导纳。

在具体应用场景中,AI 技术已深度融入周期性声子晶体、声学超材料等领域。Shi 等[79]采用 LSTM-Transformer 串联的类自编码器模型,如图 8 所示,其结合含峰值频率、空间占用等约束的多目标损失函数,实现了空间折叠声学超材料(space-folded acoustic metamaterials, SFAM)的中低频宽带隔声反设计与空间优化,在保证隔声性能达标的前提下,使结构空间占用分别降低 16.81% 和 19.39%,为空间受限场景下的声学超材料高效设计提供了可行方案,其预测性能如图 9 所示。Zea 等[80]借鉴 ResNet 架构,结合单层传声器阵列,构建从声压幅度到吸声系数间的复杂映射,

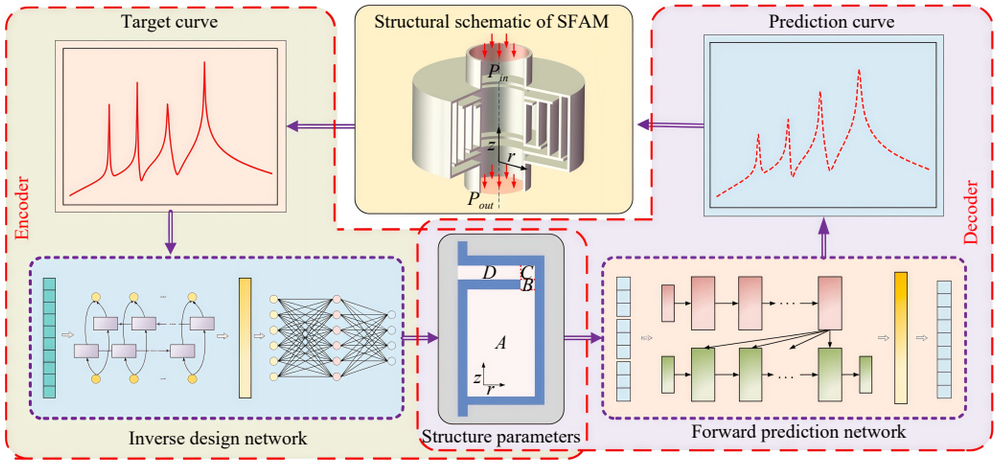



图 8 用于隔声材料设计的网络结构示意

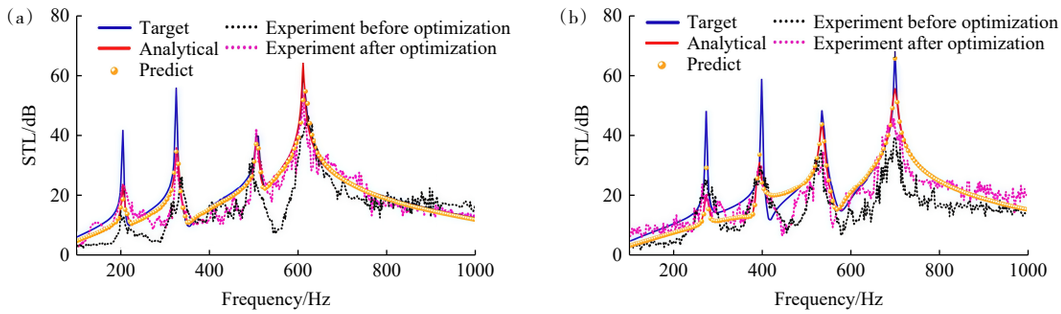



图 9 经过 AI 模型优化前后隔声量结果对比

实现宽频率范围、不同尺寸与流阻率的矩形吸声材料在强边缘衍射场景下的吸声系数的精准估计,且在400 Hz以下低频段及小尺寸吸声材料上的性能显著优于传统双麦传声器。随着AI与物理建模的融合深化,越来越多研究开始探索多物理场(如声–热、电–声)的耦合优化。例如,Wang等[81]提出的多物理信息引导PointNet(MPIPN)框架可高效解决参数声结构系统的联合优化问题。未来,基于小样本学习、自监督预训练与迁移学习的算法将进一步降低对大量仿真数据的依赖。此外,在制造可行性与实验验证方面也有显著进展,如“声学基因组工程”(phononic structure genome engineering, PSGE)平台[82]正尝试将AI设计与制造流程高度集成,实现从结构设计、性能预测到3D打印制造的闭环系统。

## 3 面临挑战

### 3.1 泛化性问题

泛化性是制约AI在声学领域落地的核心瓶颈,其本质是模型对“未见场景”的适配能力不足。Rohlfs等[83]将泛化分为样本泛化、分布泛化、域泛化、任务泛化、跨模态泛化及范围泛化6类,这里围绕这6类进行简要分析。样本泛化方面,低资源声学场景(如濒危方言识别、小众乐器分类)的标注数据量通常不足千条,且训练数据易存在标注偏差,导致模型过度拟合局部特征,数据增广技术通常需结合声学信号约束才能保证特征有效性;分布泛化的核心矛盾是测试集与训练数据的统计分布差异,如室内混响时间变化等声学环境扰动,会直接改变信号的时频分布,域适应(domain adaptation)技术通过特征分布对齐可缓解该问题,但仍需解决动态声学环境下的实时适配难题;域泛化聚焦于跨领域知识迁移,不同声学任务的特征空间存在本质差异,连续学习(continual learning)通过权重正则化缓解模型的灾难性遗忘(catastrophic forgetting),元学习(meta learning)则通过“学会学习”机制降低跨数据依赖;任务泛化面临多目标协同优化的困境,不同任务学习目标存在固有差异,可通过自监督学习与渐进式学习策略实现多任务特征的灵活

适配;跨模态泛化受限于音频、图像和文本等模态的异构性,其中 Transformer 的跨模态注意力机制已在多模态任务中验证了有效性,而最近兴起的流匹配(flow-matching, FM)等生成式技术为低资源场景下的多模态对齐提供了新路径[84];范围泛化针对参数外推能力不足,如声学仿真中训练数据有限导致的外推误差,可通过物理约束嵌入与贝叶斯不确定度(Bayesian uncertainty)估计等技术进行改善。

### 3.2 数据依赖与质量问题

数据是 AI 技术的燃料和基础,其依赖与质量问题直接制约模型的性能上限。对于声学模型而言,数据相关难题主要集中在 2 个方面:一是高质量标注数据的获取瓶颈,声学数据标注需依托专业设备(例如高精度声级计、多通道传感器阵列)与专业领域知识,流程相对复杂,成本高昂,导致低资源声学场景如濒危方言识别、特定工业噪声检测等标注数据极度稀缺,难以支撑模型充分学习通用规律;二是数据质量的固有缺陷,现有声学数据普遍存在声学环境覆盖不全、设备类型单一、样本分布不均衡等问题,且易受噪声污染、标注错误、时间标计偏差等质量影响,导致模型学习局部“伪特征”而非声学现象的本质表征。近年来,大模型的出现进一步加剧了该挑战,其对数据规模的需求呈指数级增长,需以数十万乃至亿小时级的海量音频数据为支撑,远超传统模型的数万至百万级需求,使得数据缺口被进一步放大。此外,数据依赖问题还延伸至合规性和安全层面。声学数据多涉及个人语音隐私、特定场景敏感信息,大规模数据收集极易触碰版权边界与隐私保护红线。监管政策的出台,对语料来源可追溯性、隐私安全性等提出明确要求,使得声学数据的收集流程更趋复杂。

### 3.3 复杂度问题

AI 模型的复杂度与声学应用场景的资源约束间存在突出矛盾。从模型层面看,现有通用 AI 模型的参数规模已达千亿级至万亿级,即使针对声学任务的专用模型(例如语音识别模型或语音合成模型),其参数量也多在千万至亿级,运算复杂度往往随参数规模非线性增长,导致云端部署的算力与能源成本居高不下。从应用场景看,声学技术的落地场景日益多元,助听器、全植入人工耳蜗、无线声传感网等端上设备,对模型提出了严苛的要求,通常要求 AI 模型参数量小且运算复杂度低,例如模型不超过 10 万参数,运算复杂度低于 40 MFLOPS,使得高性能模型难以适配。以芯片处理功耗水平要求仅 1 mA 的助听器为例,尽管助听器允许整个系统最大时延 10 ms,但在要求单一算法如语音增强、反馈控制(feedback control)或者宽动态范围压缩(wide dynamic range compression)仅有 2 ms 时延下,1 s 需推理 500 次,相比于 1 s 推理 1 次,同一模型推理运算复杂度相应增加 500 倍。现有解决方案可分为 3 类:一是模型压缩技术,知识蒸馏(knowledge distillation)通过“教师-学生”架构实现性能迁移;二是模型架构,如结合声学先验设计轻量级网络架构;三是硬件协同,存算一体芯片通过集成存储与运算单元,降低数据搬运带来的功耗与延时,为低功耗场景提供硬件支撑。这些技术的核心是在模型性能与复杂度之间寻求最优平衡,但在极端低资源场景下的性能损失控制仍需进一步研究。

### 3.4 实时性问题

实时性是 AI 声学技术面向实际应用的关键指标,其需求差异源于声学信号的传播特性与应用场景的功能定位。从时延要求看,不同场景的阈值跨度极大:在水声通信应用中,由于海洋中声速仅每秒 1450~1540 m[4],对于远距离通信如达到数十千米甚至上百千米,仅传播时延就达数十秒,此刻对流式处理的处理时延要求也就相对宽松。在语音关键词识别等应用中,响应时延为 300~600 ms;然而,在音视频会议系统中,则允许的时延不超过 100 ms。例如前所述的助听器或者耳机透传应用,根据最小可觉差(just noticeable difference, JND),允许的最大系统时延不超过 10 ms,否则延迟可被听觉感知,同时导致严重的梳状滤波效应。有源噪声控制(active noise cancellation, ANC)应用于耳机中,由于前馈传声器和受话器的距离可能仅 1 cm,这就要求输出时延在微秒量级才能实现主动噪声抵消。实时性的核心瓶颈包括 2 方面:一是算法复杂度,当模型运算量超过硬件算力时,实时率(real-time factor, RTF)会大于 1.0,致使处理时延失控;二是算法延迟,在声信号处理中,帧长、帧移、前看帧数以及输出数据生成方法都对算法延迟有影响,时频处理中的帧移过大或重叠相加法(overlap-add, OLA)会增加额外延迟。因而解决方案需针对性优化:针对复杂度问题,可采用轻量化模型设计、模型压

缩等技术降低运算量;针对延迟问题,可选择时域处理方法或在时频域中采用短帧移与重叠保持法(overlap-save, OLS),在助听器等场景中,可结合异步处理与滤波器时域优化将时延降至 1 个采样点级别[85]。值得注意的是,实时性往往与性能存在平衡,如何在极端时延约束下保证处理效果,是当前研究的重点方向。

### 3.5 多模态融合问题

多模态融合已成为提升声学技术性能的重要路径,但在声学领域的应用仍面临 3 类核心挑战。其一,模态异构性:音频信号的时序连续性与图像、文本的空间/离散特性存在本质差异,导致跨模态特征空间与语义对齐困难,例如视听语音增强中,视觉模态的唇部动作与音频模态的语音信号在时序同步与语义关联上难以精准匹配。其二,融合效率与性能的平衡:传统多模态融合方法(如特征拼接、加权求和等)存在信息利用率低、计算开销大等问题,难以匹配端上实时低功耗场景,而复杂的融合架构(如跨模态注意力)虽能提升性能,但会进一步加剧复杂度与实时性矛盾。其三,低资源场景的多模态数据稀疏:声学领域的多模态数据采集需同步部署多类型传感器(如声学传感器+摄像头+文本记录设备),成本高且操作不便,在低资源场景中数据稀缺,导致融合模型缺乏充分的训练基础。当前研究主要聚焦跨模态精准对齐、轻量化融合架构和低资源适配技术,未来需进一步结合声学物理规律优化,推动跨模态技术实用化。

## 4 结论与展望

AI 与声学的深度融合,推动了声学从基础研究到工程应用的全面革新。系统性梳理了 AI 在声学多个分支的应用现状,包括语音信号处理、声源定位、空间音频、声学场景与事件分类以及声学仿真与优化。通过与传统方法的对比分析,揭示了 AI 技术在性能提升、功能拓展和效率优化方面的显著优势,同时也指出当前面临的泛化性、复杂度和实时性等核心挑战。

在语音信号处理领域,AI 已实现从特征工程到端到端建模的范式转变。其中,基于深度学习的语音识别、增强和合成技术,不仅在特定任务中超越了人类水平,还通过多模态融合和生成式模型拓展了应用边界。然而,这些技术在实际部署中仍受限于数据依赖性和计算资源需求。声源定位技术得益于神经网络的强大表征能力,在复杂声学场景下的定位精度和鲁棒性显著提升,但跨场景和实时处理能力仍需进一步优化。空间音频技术通过 AI 技术实现了声场重构、个性化 HRTF 预测等技术突破,但受限于人类听觉机理的复杂性和高质量数据的稀缺,其性能与传统方法相比尚未形成全面优势。声学场景与事件分类技术借助预训练与大规模数据集,展现了强大的识别与多任务适应能力,但在小样本中的表现仍有一定提升空间。声学仿真与优化领域通过 AI 驱动的逆向设计,大幅提升了结构优化的效率,但物理约束嵌入与多目标协同优化的平衡仍需进一步探索。

AI 在声学中的应用将呈现以下发展趋势:首先,跨模态与多任务协同将成为技术突破的关键方向。声学问题往往涉及声音与视觉等的信息交互,未来研究需进一步探索多模态信息的深度融合机制;同时,多任务学习框架的进一步优化将解决声学任务中多目标冲突的问题,实现资源共享与性能的平衡。其次,小样本与自监督技术将缓解强数据依赖问题,当前 AI 模型严重依赖数据标注,而在低资源场景中,数据获取成本相对高昂。未来,基于自监督学习和元学习的框架可以广泛应用于声学任务中,通过挖掘数据内在规律和跨领域知识迁移,降低算法对数据的依赖。此外,物理信息引导的生成式模型有望生成更符合真实声学规律的数据,进一步提升模型泛化能力;再者,边缘计算与轻量化部署将推动基于 AI 的声学技术的普及。随着芯片技术的不断进步与模型压缩技术的成熟将逐步推动复杂模型在助听器、物联网等资源受限场景中的高效运行。存算一体化芯片和专用硬件加速器的应用,将显著降低功耗与延时,满足实时性要求极高的应用(如主动降噪、助听器等)的需求。最后,物理机理与数据驱动的深度融合将提升模型的可靠性与可解释性。声学作为物理学科的分支,其本质规律应被更紧密地嵌入 AI 技术中。通过结合物理信息规律,不仅能提升外推预测的准确性,还能为声学设计提供可解释性的指导原则。

AI 在声学中的应用和发展也面临诸多挑战。在基础理论层面,声学与 AI 的交叉研究尚未建立完善

的理论框架,需要重点研究以指导模型的设计和性能评估。在技术层面,如何平衡算法复杂度与性能,实现可扩展性的实时处理,仍是亟待解决的关键技术难题。此外,在伦理与隐私保护方面也需要行业规范和技术防护的双重保障。

AI 与声学的结合正处于快速发展阶段,其潜力尚未充分释放。未来,随着基础理论的突破、技术的迭代和跨学科合作的深化,“AI+声学”将在海洋探测、医疗诊断、虚拟现实、环境声学等领域进一步发挥重要的作用,以最终实现从实验室研究、单点技术落地到大规模产业化应用的跨越。

## 参考文献(References)


[1] Morse P M, Ingard K U. Theoretical Acoustics[M]. New Jersey: Princeton University Press, 1987.

[2] 马大猷. 现代声学理论基础[M]. 北京: 科学出版社, 2004.

[3] 程建春, 李晓东, 杨军. 声学学科现状以及未来发展趋势[M]. 北京: 科学出版社, 2021.

[4] 汪德昭. 水声学[M]. 北京: 科学出版社, 1981.

[5] 应崇福. 超声学[M]. 北京: 科学出版社, 1990.

[6] Howard D M, Angus J. Acoustics and psychoacoustics[M]. 5th Edition. New York: Routledge, 2017.

[7] 蔡自兴, 蔡昱峰. 人工智能学科体系的构建与探讨[J]. 科技导报, 2025, 43(8): 15−26.

[8] 山世光, 阚美娜, 刘昕, 等. 深度学习: 多层神经网络的复兴与变革[J]. 科技导报, 2016, 34(14): 60−70.

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

[10] Amodei D, Ananthanarayanan S, Anubhai R, et al. Deep speech 2: End−to−end speech recognition in English and mandarin[C]//Proceedings of the 33rd International Conference on Machine Learning (ICML). New York: PMLR, 2016: 173−182.

[11] Xiong W, Wu L, Alleva F, et al. The microsoft 2017 conversational speech recognition system[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5934−5938.

[12] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity[J]. Bulletin of Mathematical Biology, 1943, 5(4): 115−133.

[13] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain[J]. Psychological Review, 1958, 65(6): 386−408.

[14] Rumelhart D E, Hinton G E, Williams R J. Learning internal representations by error propagation[J]. Nature, 1986, 323: 533−536.

[15] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer−wise training of deep networks[C]//Proceedings of Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. MIT PressSept. 2007: 153−160.

[16] LeCun Y, Bottou L, Bengio Y, et al. Gradient−based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278−2324.

[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems (NIPS). Cambridge MA: MIT Press, 2012: 1097−1105.

[18] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV: IEEE, 2016: 770−778.

[19] Hochreiter S, Schmidhuber J. Long short−term memory[J]. Neural Computation, 1997, 9(8): 1735−1780.

[20] Cho K, van Merrienboer B, Bahdanau D, et al. On the properties of neural machine translation: Encoder–decoder approaches[C]//Proceedings of SSST−8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Doha, Qatar: Association for Computational Linguistics (ACL), 2014: 103−111.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems (NIPS). Long Beach, California: MIT Press, 2017: 1−11.

[22] Zheng C S, Zhang H Y, Liu W Z, et al. Sixty years of frequency−domain monaural speech enhancement: From traditional to deep learning methods[J]. Trends in Hearing, 2023, 27: 23312165231209913.

[23] Radford A, Kim J W, Xu T, et al. Robust speech recognition via large−scale weak supervision[C]//Proceedings of the 40th International Conference on Machine Learning (ICML). Honolulu, HI: PMLR, 2023: 28492−28518.

[24] Afouras T, Chung J S, Senior A, et al. Deep audio−visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(12): 8717−8727.

[25] Kamble A, Ghare P H, Kumar V, et al. Spectral analysis of EEG signals for automatic imagined speech recognition[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 4009409.

[26] Wang D L, Chen J T. Supervised speech separation based on

deep learning: An overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702−1726.

[27] Xu Y, Du J, Dai L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 7−19.

[28] Pandey A, Wang D L. A new framework for CNN−based speech enhancement in the time domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179−1188.

[29] Richter J, Welker S, Lemercier J M, et al. Speech enhancement and dereverberation with diffusion−based generative models[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2351−2364.

[30] Michelsanti D, Tan Z H, Zhang S X, et al. An overview of deep−learning−based audio−visual speech enhancement and separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1368−1396.

[31] Kawahara H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds[J]. Acoustical Science and Technology, 2006, 27(6): 349−353.

[32] van den Oord A, Dieleman S, Zen H, et al. WaveNet: A generative model for raw audio[EB/OL]. (2016−09−12) [2025−11−20]. https://arxiv.org/abs/1609.03499.

[33] Ren Y, Ruan Y, Tan X, et al. Fastspeech: Fast, robust and controllable text to speech[C]// Advances in Neural Information Processing Systems (NeurIPS). Vancouver, Canada: MIT Press, 2019: 1−10.

[34] Chen S Y, Wang C Y, Wu Y, et al. Neural codec language models are zero−shot text to speech synthesizers[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025, 33: 705−718.

[35] Wightman F L, Kistler D J. Monaural sound localization revisited[J]. The Journal of the Acoustical Society of America, 1997, 101(2): 1050−1063.

[36] Sun X C, Jia H, Zhang Z, et al. Sound localization and separation in 3D space using a single microphone with a metamaterial enclosure[J]. Advanced Science, 2020, 7(3): 1902271.

[37] Blauert J. Spatial Hearing: The psychophysics of human sound localization[M]. Revised Version. Cambridge, MA: MIT Press, 1996.

[38] Fowlkes J B. From sonar to medical ultrasound—The impact of Paul Langevin[J]. The Journal of the Acoustical Society of America, 2022, 152(4): A30.

[39] Richardson L F. Apparatus for warning a ship of its approach to large objects in fog[P]. British: British Patent, 1913.

[40] Duck F. Ultrasound−the first fifty years[J]. Medical Physics International Journal, 2021(5): 470−798.

[41] Wax M, Kailath T. Optimum localization of multiple sources by passive arrays[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1983, 31(5): 1210−1217.

[42] Schmidt S. A signal subspace approach to multiple emitter localization and spectral estimation[D]. Stanford, CA, USA: Stanford University, 1981.

[43] Knapp C, Carter G. The generalized correlation method for estimation of time delay[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1976, 24(4): 320−327.

[44] Donohue K D, Hannemann J, Dietz H G. Performance of phase transform for detecting sound sources with microphone arrays in reverberant and noisy environments[J]. Signal Processing, 2007, 87(7): 1677−1691.

[45] Jekateryńczuk G, Piotrowski Z. A survey of sound source localization and detection methods and their applications[J]. Sensors, 2023, 24(1): 68.

[46] Grumiaux P A, Kitić S, Girin L, et al. A survey of sound source localization with deep learning methods[J]. The Journal of the Acoustical Society of America, 2022, 152(1): 107−151.

[47] Krause D, Politis A, Kowalczyk K. Feature overview for joint modeling of sound event detection and localization using a microphone array[C]//Proceedings of 28th European Signal Processing Conference (EUSIPCO). Amsterdam, The Netherlands: IEEE, 2020: 31−35.

[48] Zhang G, Geng L, Xie F, et al. A dynamic convolution−transformer neural network for multiple sound source localization based on functional beamforming[J]. Mechanical Systems and Signal Processing, 2024, 211: 111272.

[49] Diaz−Guerra D, Politis A, Virtanen T. Position tracking of a varying number of sound sources with sliding permutation invariant training[C]//Proceedings of 31st European Signal Processing Conference (EUSIPCO). Helsinki, Finland: IEEE, 2023: 251−255.

[50] Wang Q, Du J, Wu H X, et al. A four−stage data augmentation approach to ResNet−conformer based acoustic modeling for sound event localization and detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1251−1264.

[51] Bianco M J, Gannot S, Gerstoft P. Semi−supervised source localization with deep generative modeling[C]//Proceedings of IEEE 30th International Workshop on Machine Learning

for Signal Processing (MLSP). Espoo, Finland: IEEE, 2020: 1−6.

[52] Chakrabarty S, Habets E A P. Multi−speaker DOA estimation using deep convolutional networks trained with noise signals[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(1): 8−21.

[53] Cobos M, Ahrens J, Kowalczyk K, et al. An overview of machine learning and other data−based methods for spatial audio capture, processing, and reproduction[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2022, 2022(1): 10.

[54] 谢菠荪. 空间声原理及其相关的物理、听觉问题[J]. 中国科学: 物理学 力学 天文学, 2022, 52(4): 244−303.

[55] Lluís F, Martínez−Nuevo P, Bo Mller M, et al. Sound field reconstruction in rooms: Inpainting meets super−resolution[J]. The Journal of the Acoustical Society of America, 2020, 148(2): 649−659.

[56] Ma F, Zhao S P, Burnett I S. Sound field reconstruction using a compact acoustics−informed neural network[J]. The Journal of the Acoustical Society of America, 2024, 156(3): 2009−2021.

[57] Park S Y, Chun C J, Kim H K. Subband−based upmixing of stereo to 5.1−channel audio signals using deep neural networks[C]//Proceedings of International Conference on Information and Communication Technology Convergence (ICTC). Jeju, South Korea: IEEE, 2016: 377−380.

[58] Choi J, Chang J H. Exploiting deep neural networks for two−to−five channel surround decoder[J]. Journal of the Audio Engineering Society, 2021, 68(12): 938−949.

[59] Lee G W, Kim H K. Personalized HRTF modeling based on deep neural network using anthropometric measurements and images of the ear[J]. Applied Sciences, 2018, 8(11): 2180.

[60] Yao D D, Zhao J L, Cheng L B, et al. An individualization approach for head−related transfer function in arbitrary directions based on deep learning[J]. JASA Express Letters, 2022, 2(6): 064401.

[61] Gao R H, Grauman K. 2.5D visual sound[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 324−333.

[62] Morgado P, Li Y, Nvasconcelos N. Learning representations from audio−visual spatial alignment[C]//Advances in Neural Information Processing Systems (NeurIPS), Vancouver, Canada: MIT Press, 2020: 1−11.

[63] 毛玉如, 汪赟, 张建勋. 中国噪声自动监测的现状和发展[J]. 科技导报, 2024, 42(20): 23−31.

[64] 袁旻忞, 王彦琴, 邵社刚, 等. 道路交通噪声控制技术研发进展[J]. 科技导报, 2024, 42(20): 85−94.

[65] 户文成, 刘磊, 康钟绪, 等. 城市轨道交通噪声污染防治进展[J]. 科技导报, 2024, 42(20): 70−84.

[66] Mesaros A, Heittola T, Virtanen T, et al. Sound event detection: A tutorial[J]. IEEE Signal Processing Magazine, 2021, 38(5): 67−83.

[67] Mesaros A, Heittola T, Virtanen T. TUT database for acoustic scene classification and sound event detection[C]//Proceedings of 24th European Signal Processing Conference (EUSIPCO). Budapest, Hungary: IEEE, 2016: 1128−1132.

[68] Jallet H, Cakır E, Virtanen T. Acoustic scene classification using convolutional recurrent neural networks[C]//Detection and Classification of Acoustic Scenes and Events(DCASE), Munich, Germany: IEEE, 2017: 1−3.

[69] Gong Y, Chung Y A, Glass J. AST: Audio spectrogram transformer[C]//Proceedings of Interspeech 2021. Brno, Czech Republic: ISCA, 2021: 571−575.

[70] Gemmeke J F, Ellis D P W, Freedman D, et al. Audio Set: An ontology and human−labeled dataset for audio events[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA, USA: IEEE, 2017: 776−780.

[71] Kong Q Q, Cao Y, Iqbal T, et al. PANNs: Large−scale pretrained audio neural networks for audio pattern recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880−2894.

[72] Han B, Jiang A B, Zheng X H, et al. Exploring self−supervised audio models for generalized anomalous sound detection[J]. IEEE Transactions on Audio, Speech and Language Processing, 2025, 33: 4126−4141.

[73] 杭银, 李秋彤, 刘艳, 等. 声学超材料中的智能设计: 从优化到逆向设计[J]. 功能材料与器件学报, 2025, 31(2): 88−98.

[74] Dühring M B, Jensen J S, Sigmund O. Acoustic design by topology optimization[J]. Journal of Sound and Vibration, 2008, 317(3/4/5): 557−575.

[75] Donda K, Brahmkhatri P, Zhu Y F, et al. Machine learning for inverse design of acoustic and elastic metamaterials[J]. Current Opinion in Solid State and Materials Science, 2025, 35: 101218.

[76] Luo C C, Ning S W, Liu Z L, et al. Interactive inverse design of layered phononic crystals based on reinforcement learning[J]. Extreme Mechanics Letters, 2020, 36: 100651.

[77] Raissi M, Perdikaris P, Karniadakis G E. Physics−informed neural networks: A deep learning framework for solving

forward and inverse problems involving nonlinear partial differential equations[J]. Journal of Computational Physics, 2019, 378: 686−707.

[78] Schmid J D, Bauerschmidt P, Gurbuz C, et al. Physics−informed neural networks for acoustic boundary admittance estimation[J]. Mechanical Systems and Signal Processing, 2024, 215: 111405.

[79] Shi Z H, Chen C Z, Zhang D C, et al. Inverse design and spatial optimization of SFAM via deep learning[J]. International Journal of Mechanical Sciences, 2025, 306: 110855.

[80] Zea E, Brando E, Nolan M, et al. Sound absorption estimation of finite porous samples with deep residual learning [J]. The Journal of the Acoustical Society of America, 2023, 154(4): 2321−2332.

[81] Wang C, Wu J H, Wang Y Z, et al. MPIPN: A multi physics−informed PointNet for solving parametric acoustic−structure systems[J]. Engineering with Computers, 2025, 41(1): 225−246.

[82] Dong H W, Shen C, Liu Z, et al. Inverse design of phononic meta−structured materials[J]. Materials Today, 2024, 80: 824−855.

[83] Rohlfs C. Generalization in neural networks: A broad survey[J]. Neurocomputing, 2025, 611: 128701.

[84] Li S F, Kallidromitis K, Gokul A, et al. OmniFlow: Any−to−any generation with multi−modal rectified flows[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2025: 13178−13188.

[85] Zheng C S, Liu W Z, Li A D, et al. Low−latency monaural speech enhancement with deep filter−bank equalizer[J]. The Journal of the Acoustical Society of America, 2022, 151(5): 3291−3304.

# Applications and prospects of artificial intelligence in acoustics

ZHENG Chengshi[1,2], LI Andong[1,2], RAO Dan[3], YUAN Minmin[4], JIANG Feng[1], LI Xiaodong[1,2]

1. Laboratory of Noise and Audio Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. School of Physics and Optoelectronics, South China University of Technology, Guangzhou 510641, China
4. Research Institute of Highway, Minisry of Transport, Beijing 100088, China

**Abstract** Artificial intelligence is regarded as the core driving force of the Fourth Industrial Revolution. While it brings impacts and challenges to many traditional industries, it also plays an important role in empowering and enhancing their quality. Similarly, it has brought new development opportunities to the ancient discipline of acoustics. Currently, artificial intelligence has been deeply intersecting and integrating with underwater acoustics, ultrasonics, and air acoustics, continuously promoting the innovation of acoustic technologies. This paper focuses on the applications of artificial intelligence in acoustics, especially in air acoustics. Firstly, we elaborate in detail on the current applications of artificial intelligence in many fields, such as speech signal processing, sound source localization, spatial audio, acoustic event detection, classification and monitoring, as well as acoustic simulation and optimization, and we then further analyze the advantages of using artificial intelligence when compared with traditional methods. Secondly, in terms of the possible core problems which can hinder applications in practical scenarios, we conduct comprehensive discussions, including generalization, data dependency and low quality, computational complexity, real−time deployment, and multi−modality fusion. Finally, we summarize the challenges faced by the application of artificial intelligence in acoustics and the future development directions.

**Keywords** artificial intelligence; deep learning; acoustics; audio signal processing; sound source localization; spatial audio

（责任编辑 徐丽娇）