Sound Source Localization and Tracking Based on Deep Learning: A Survey

CHEN Zhe, SONG Dengao, WANG Yiyu, YIN Fuliang*

(School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: Sound source localization and tracking constitute an important means for machine hearing to acquire spatial information. With the growing adoption of multi-microphone devices in applications such as speech interaction, conference systems, and acoustic monitoring, the demand for stable estimation of a sound source's direction and position in complex acoustic environments continues to increase. Accordingly, this paper presents a systematic review of deep-learning-based techniques for sound source localization and tracking. Existing review articles have mainly focused on sound source localization, whereas deep-learning-based sound source tracking has not yet been systematically reviewed. To fill this gap, this paper presents a unified analysis of both sound source localization and tracking. First, the fundamental problem formulation and the framework of traditional approaches are outlined. Then, from the perspectives of input representation, model architecture, and learning objectives, the main lines of deep learning methods are introduced with respect to feature design, network modeling, and training strategies. Next, commonly used datasets, experimental settings, and evaluation metrics are summarized, and key considerations for comparing results under different conditions are discussed. Finally, the reviewed techniques are summarized and potential future research directions are outlined.

Highlights:

1. This paper systematically reviews research on deep learning-based sound source localization and tracking, with particular emphasis on the technological evolution from instantaneous spatial localization to continuous trajectory estimation.
2. The development of mainstream methods is summarized from the perspectives of input representation, network architecture, and temporal modeling, covering typical deep learning models such as CNN, RNN/LSTM, CRNN, and Transformer.
3. The performance advantages of deep learning-based methods in noisy, reverberant, multi-source-overlapping, and dynamic scenarios are summarized, and future directions are identified, including robustness in real-world environments, generalization ability, and lightweight deployment.

Key words: sound source localization; sound source tracking; neural networks; deep learning

---

Foundation items: National Natural Science Foundation of China (Nos.62271103, 61871066).

Received: 2026-01-09; **Revised**: 2026-02-26

***Corresponding author**, **E-mail**: flyin@dlut.edu.cn.

ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 2,Mar. 2026,pp. 371－396
DOI:10. 16337/j. 1004-9037. 2026. 02. 006

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

基于深度学习的声源定位与跟踪综述

陈　喆，宋登鏊，王一宇，殷福亮

（大连理工大学信息与通信工程学院，大连 116024）

摘　要：声源定位与跟踪是机器听觉获取空间信息的重要途径之一。随着多麦克风设备与语音交互、会议系统和声学监测等应用的发展，在复杂声场条件下对声源方向与位置进行稳定估计的需求持续增加。基于此，本文对基于深度学习的声源定位与跟踪相关技术进行了系统综述。现有综述多聚焦于声源定位，而对基于深度学习的声源跟踪研究缺乏系统梳理。针对这一不足，本文将声源定位与跟踪纳入统一框架进行综合分析。首先，概述了声源定位与跟踪的基本问题定义与传统方法框架。然后，从输入表征、模型结构与学习目标三个角度，介绍了深度学习方法在特征设计、网络建模以及训练策略方面的主要路线。接着，总结了常用数据集、实验设置与评价指标，并讨论不同条件下结果对比的注意事项。最后，对声源定位与跟踪技术进行总结，并对未来可能的研究方向进行展望。

关键词：声源定位；声源跟踪；神经网络；深度学习

中图分类号：TN912.3　　**文献标志码：**A

引用格式：陈喆，宋登鏊，王一宇，等. 基于深度学习的声源定位与跟踪综述[J]. 数据采集与处理，2026，41(2)：371-396. CHEN Zhe，SONG Dengao，WANG Yiyu，et al. Sound source localization and tracking based on deep learning：A survey[J]. Journal of Data Acquisition and Processing，2026，41(2)：371-396.

引　　言

随着野生动物监测[1]、安防监控[2]，机器人技术[3-4]和3D场景感知[5]等领域的快速发展，对空间声源信号的高效处理与分析需求日益提升。声源定位与跟踪作为空间声源检测的重要技术之一，主要通过多通道麦克风阵列观测估计声源在空间中的方向或位置，并进一步在时间维度上保持其轨迹的连续性与目标身份的一致性，从而有效实现对发声个体的定位与运动轨迹跟踪。具体而言，声源定位（Sound source localization，SSL）侧重于在单帧内输出声源的方位角与俯仰角或三维坐标，而跟踪则是在连续帧序列上融合历史信息与运动先验，实现对动态声源的时序滤波、轨迹平滑以及多目标情况下的目标管理与数据关联，其根本目的在于为机器提供稳定、可用且可解释的空间听觉能力，从而支撑更高层的感知与决策任务。传统声源定位与跟踪方法通常依赖显式声学模型与手工构造的空间特征，例如基于互相关的时延估计[6]、基于空间谱扫描的波束形成与功率响应[7-8]，以及基于子空间分解的高分辨率估计[9]，并在跟踪阶段配合卡尔曼滤波、粒子滤波与多目标数据关联等策略完成轨迹更新。

虽然声源定位与跟踪已广泛应用于真实应用中，但还存在复杂混响与多径传播导致测量分布多峰化与虚假峰增强[10-11]的问题，同时还会出现多源重叠与非平稳噪声引起的观测缺失与关联歧义[11-12]，另外在真实环境中阵列几何变化、通道失配与同步误差也会又进一步破坏模型假设[13-14]，扫描式或高阶统计估计带来的计算开销也使在线部署受到约束[15-16]，这都使得传统方法在鲁棒性与泛化性上面临瓶颈。

基金项目：国家自然科学基金(62271103，61871066)。
收稿日期：2026-01-09；**修订日期：**2026-02-26

随着深度学习的发展,数据驱动方法作为传统模型驱动方案的补充,能够在保留物理可解释约束的同时,通过深度网络学习空间线索的提取与融合,结合不确定性表达与时序建模,改善多径、噪声与多源场景下的定位与跟踪性能。

近年来基于深度学习的声源定位与跟踪系统数量不断增加[17],大多数已发表的研究表明,基于深度学习的声源定位与跟踪方法优于传统方法。目前研究逐步从“逐帧空间估计”走向“时序一致的轨迹推断”,并在表征、结构与学习目标3个层面持续演进。早期工作多以通道间互相关、相位差等空间线索为核心输入[18-19],学习从多通道观测到方向或位置的映射关系;随着研究深入,输入表征进一步扩展到多尺度时频表示以及更具阵列可迁移性的空间表征,端到端波形建模也被用于减少对手工特征与理想假设的依赖[20],从而提升复杂声场下的鲁棒性[21]。

在建模层面,卷积网络为定位提供了有效的空间判别能力,随后与循环单元或时序卷积相结合,使模型能够显式利用跨帧上下文以维持输出的时间连续性,并形成在联合定位与检测任务中广泛采用的基线范式[22-24]。进一步地,自注意力机制被引入以强化长程依赖建模与跨通道全局交互[25],改善多源共存、强混响以及空间线索随时间快速变化时的稳定性。与此同时,音频与视觉等多模态信息的融合为遮挡与低信噪比条件下的持续跟踪提供了额外约束[26],弱监督与自监督学习则在一定程度上缓解了精确空间标注成本高、跨场景泛化困难等问题[27-28]。尽管上述方向已积累了大量研究成果,但现有综述多数仍以定位或定位检测为主线[17],对定位输出向轨迹形成的关键机制缺少统一梳理,尤其是在不确定性表达、跨帧一致性约束、多目标关联与变源数管理,以及端到端时序建模与显式跟踪器协同边界等方面尚未形成系统化的比较框架,本文旨在围绕这些核心环节对相关工作进行结构化整理与归纳。

基于深度学习的声源定位与跟踪系统的核心处理逻辑可概括为如图1所示的级联架构。通过麦克风阵列采集的多通道输入信号首先经过特征提取模块,转化为包含空间线索的高维声学特征,这些特征随后被送入深度学习定位模型,通过非线性映射解算出当前时刻瞬时的位置信息。与单帧定位不同,为了解决环境噪声干扰及声源移动带来的不确定性,定位模型的输出并不直接作为最终结果,而是作为观测数据传递给跟踪模型,该模块利用时序上下文对声源状态进行滤波与平滑,最终输出连续、鲁棒的坐标轨迹预测。这种“先定位、后跟踪”的分层设计是当前解决动态声源分析任务的主流范式[29]。

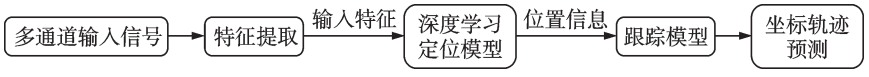

图1 基于深度学习的声源定位与跟踪的通用流程

Fig.1 General pipeline for deep-learning-based sound source localization and tracking

本文旨在系统梳理近年来基于深度学习的声源定位与跟踪技术进展,围绕“声学观测-空间参数”之间的非线性映射学习,归纳深度模型在复杂噪声、混响及多声源条件下的鲁棒性提升机制。重点从网络结构视角对深度前馈网络(Deep neural network, DNN)、卷积神经网络(Convolutional neural network, CNN)、循环神经网络(Recurrent neural network, RNN)及其长短期记忆(Long short-term memory, LSTM)变体,以及近年来引入的自注意力Transformer等典型架构进行系统比较,分析其在空间特征建模、时序信息利用与复杂声场适应性方面的优势与局限,并进一步讨论端到端建模、多源可扩展输出与多模态融合等方向对定位与跟踪性能的促进作用。

1 基本问题定义与方法框架

1.1 声源定位与跟踪的问题定义及基本建模设定

本文重点关注室内场景中的声源定位与跟踪问题,即麦克风阵列在固定或缓慢变化的空间构型下

采集多通道信号,声源可为静态或运动目标,且真实观测普遍受到混响、多径、噪声、干扰以及间歇发声等因素影响。在此类条件下,定位任务通常以到达方向(Direction of arrival, DOA)或空间位置作为输出,而跟踪任务进一步要求在时间轴上维持输出的连续性与一致性,并在多源情况下处理目标出现与消失、重叠发声导致的观测缺失以及由此引发的关联歧义。该问题设定与近年公开评测框架对“在干扰与噪声条件下识别声事件的活动并估计其空间方向或位置”的定义一致,也与LOCATA等面向真实录音的基准对动态几何变化、漏检与误检等工程挑战的描述相吻合[11]。

在信号与阵列层面,假设多通道信号在短时窗内近似分段平稳,这一假设允许基于时频分析构造输入表征或直接端到端学习。对部分方法而言,空间几何关系可在频带内通过时延或相位差近似刻画,但在宽带语音与复杂室内声场中,这类近似往往难以完整涵盖近远场切换、频率相关传播以及强混响引起的多径叠加,因此深度学习方法通常通过学习式表征来吸收模型失配带来的偏差,并以更鲁棒的方式从多通道观测中提取空间线索。

在采集条件方面,一方面,假定阵列几何在训练与推理阶段可重复或可标定,即麦克风相对位置与朝向已知,通道同步在工程允许范围内成立,且增益与频响差异已完成校准,或由模型在训练中隐式吸收。在此类设定下,文献中最常见的输入通道规模为四通道阵列,这一选择与公开基准和评测框架的长期配置密切相关,例如DCASE SELD系列数据集通常以两种包含四通道空间信息的三维录制格式作为输入,即一阶全景声与四面体麦克风阵列[30-31]。另一方面,面向可穿戴与消费级设备的研究也经常采用双通道或近双通道配置,以匹配硬件条件并考察在受限阵列孔径下的可辨识性与鲁棒性。此外,在机器人听觉与高精度阵列声学设置中也存在更高通道数的配置,例如LOCATA基准[11]包含每侧两麦的助听器配置、集成于机器人头部的12麦阵列以及32麦球阵,常用于评估复杂真实录音条件下的定位与跟踪能力,并反映高通道阵列在精度与稳健性上的潜在优势。当阵列几何、通道特性或采集设备发生变化时,深度模型可能出现显著的域偏移,因此阵列不敏感表示、跨设备泛化与域适配是近年来被持续关注的研究方向之一。

基于上述前提,本文所综述方法的适用范围覆盖单源与多源定位、静态与运动声源跟踪,以及以语音与通用声事件为代表的两类典型目标配置。对于多源与变源数场景,本文默认算法需面对观测的多峰性与间断性,并对轨迹管理与关联策略提出要求,后续章节将结合输入表示、模型结构对其进行专门讨论。

1.2　传统声源定位与跟踪方法

在深度学习方法兴起之前,研究者已发展出一系列基于阵列信号处理的技术用于声源定位与跟踪,这些传统方法不仅常被作为深度学习方法的对比基线,也经常为后续学习模型提供由经典算法构造的空间特征与先验信息。因此,在进入深度学习方法谱系之前,有必要先回顾传统范式的基本思路与处理链路:通常先对多通道信号进行时频分析并提取通道间空间线索(如互相关时延或相位差),据此形成逐帧的位置观测,再结合贝叶斯滤波与数据关联实现跨帧平滑、轨迹更新以及多目标管理。本节将据此对最常见的传统方法作简要介绍。

1.2.1　*传统声源定位方法*

在传统阵列声源定位中,研究通常先构造能够反映空间信息的“观测”,再据此推断声源的到达方向或空间位置。围绕观测的构造方式与推断机制,经典方法大体可以归纳为3条主线:以时延或相位一致性为核心的空间观测方法、以二阶统计量为基础的谱估计方法和以生成模型或源分离假设为基础的统计建模方法。这样的划分既覆盖最常用的算法家族,也便于后续讨论深度学习如何学习这些空间线索、如何替代其中的部分环节,以及在多源与移动场景中需要哪些额外机制来维持输出的时序稳定性。

第1条主线以通道间时间差或相位差作为空间线索。在阵列几何已知时,传统定位通常先估计麦

克风对的到达时间差(Time difference of arrival, TDOA),再由几何约束推断声源到达方向或空间位置。单源场景中,常用的相位变换广义互相关(Generalized cross-correlation with phase transform, GCC-PHAT)[6]通过对互功率谱进行相位归一化以增强直达声相关峰,从而获得更稳健的时延估计。对移动声源,可将瞬时DOA作为观测并结合卡尔曼滤波等实现轨迹平滑与遮挡鲁棒性。在多源或强混响条件下,单对麦克风的时延峰易出现多峰与漂移,因此常将一致性扩展到阵列层面,通过空间扫描形成能量一致性图。相位变换导向响应功率(Steered response power with phase transform, SRP-PHAT)[32]对候选方向或网格计算理论时延并累加对应相关值以形成空间谱,并以峰值作为方向或位置估计。该方法实现直观且对幅度失配相对不敏感,但依赖网格搜索,分辨率提高会显著增加开销,并在强混响或源重叠时易出现非目标峰与漂移;移动场景中峰值可能交叉,逐帧取峰难以形成稳定轨迹,通常需要时序平滑与后端关联共同维护。

第2条主线基于二阶统计量,利用多通道协方差矩阵结构构造高分辨率空间谱。该类方法将短时统计结构作为关键载体,通过矩阵分解或约束优化分离与方向相关的子结构。多重信号分类(Multiple signal classification, MUSIC)[33]通过协方差特征分解分离信号与噪声子空间并构造尖峰伪谱,具有较高角分辨率与多源可分性,但对标定误差、快拍不足及相干源更敏感,工程上常用宽带融合、对角加载或空间平滑增强稳健性。移动多源场景下需要权衡窗口长度,窗口过短估计不稳,过长会弱化动态变化。在强干扰背景下,自适应波束形成也常用于定位。最小方差无失真响应(Minimum variance distortionless response, MVDR)[34]在无失真约束下最小化输出功率,通常能获得更尖锐的谱峰并抑制旁瓣干扰,但对混响、多径、模型失配及协方差估计误差更敏感,并依赖足够快拍保证矩阵估计稳定。移动场景中同样存在窗口与非平稳性矛盾,使得逐帧谱峰往往仍需与时间建模和后端跟踪协同。

第3条主线侧重统计生成与源分离建模,将多源观测视为多个潜在成分的联合解释。该类方法的核心逻辑在于利用统计推断技术从混叠信号中解构出各声源的独立贡献,从而在概率框架下处理多源干扰与重叠问题。具体而言,独立成分分析与盲源分离[35]基于近似独立性分解混合信号并结合空间一致性推断方向;混合模型方法如高斯混合模型与高斯混合回归[36]利用时频稀疏性,将空间特征建模为多方向混合并通过期望最大化估计软归属。该类方法解释性较强,可缓解硬判决歧义,但依赖统计假设与参数估计稳定性,计算开销也可能限制在线应用,因此通常仍需与时序平滑、轨迹管理与关联策略结合以获得连续可靠的跟踪结果。

1.2.2 传统声源跟踪方法

在上一节中,得到的是逐帧的位置估计,传统跟踪模块的核心作用是把逐帧定位结果视作观测序列,在引入运动学约束的同时进行递归融合,从而获得更平滑连续的轨迹估计。这一思想在贝叶斯滤波框架下可以自然表述为“状态预测+观测校正”,而卡尔曼滤波正是其中在线性高斯假设下的经典实现。

卡尔曼滤波(Kalman filter, KF)[37]由Kalman于1960年提出,是一种递归的贝叶斯最优状态估计方法,在目标跟踪与信号处理等任务中应用广泛。其核心是在状态空间框架下融合运动模型与观测信息,对线性高斯系统给出最优的线性无偏估计。在声源定位与跟踪中,卡尔曼滤波可理解为将声源运动状态与定位观测统一到同一递推模型中:将声源的位置与速度等记为状态向量,将定位模块输出的方向或到达时间差等记为观测向量,并用状态转移方程与观测方程描述二者关系。在线性高斯假设下,系统可写为线性状态方程与线性观测方程,其中过程噪声与观测噪声通常建模为零均值高斯分布,协方差分别为过程噪声协方差与观测噪声协方差。滤波递推由预测与更新两阶段组成:预测阶段利用状态模型推进状态及其误差协方差,更新阶段根据观测残差计算卡尔曼增益并修正状态估计与协方差,从而实现对轨迹的连续平滑估计。

由于声源定位任务中观测量与空间状态通常非线性映射,若仍采用线性观测模型易产生模型失配并降低跟踪性能,因此常引入扩展卡尔曼滤波(Extended KF,EKF)[38]处理非线性情形。EKF将状态方程与观测方程推广为非线性函数,并在当前估计点附近对非线性函数做一阶泰勒展开,以相应的雅可比矩阵近似线性模型中的状态转移矩阵与观测矩阵。其递推结构仍保持预测与更新两阶段:预测阶段用线性化后的状态模型传播先验状态与协方差,更新阶段用线性化后的观测模型计算增益并基于观测残差校正先验估计,循环迭代即可得到对移动声源状态的连续估计。实际系统中通常以位置与速度作为状态,以方向估计或到达时间差估计作为观测,通过融合运动学模型与定位观测实现轨迹的稳定跟踪。

1.3　基于深度学习的声源定位与跟踪方法框架

与传统方法主要依赖显式声学模型、阵列几何关系以及人工设计空间特征不同,基于深度学习的声源定位与跟踪方法更多采用数据驱动的建模思路,直接学习多通道观测信号与声源空间状态之间的映射关系。总体来看,这类方法通常围绕输入表征、特征建模、输出预测和后处理等环节展开,并在统一框架下完成空间信息的提取、建模与估计。

在输入端,模型通常以多通道音频信号为基础,将其表征为通道间特征、互相关特征、频谱图特征、高阶环境声(Ambisonics)特征与声强特征、原始波形特征以及其他补充性特征等不同输入形式,从而尽可能保留与声源方向、位置及运动状态相关的空间线索。在此基础上,模型进一步通过深度神经网络对多通道观测中的空间相关性与时间连续性进行建模。常见的网络结构包括CNN、RNN、LSTM以及Transformer等,这些模型能够从复杂声场中的多通道信号中提取更具判别性的空间特征表示,并增强系统在噪声、混响及多源条件下的建模能力。

从输出形式来看,基于深度学习的方法会根据具体任务目标给出不同类型的预测结果,例如声源的到达方向、空间位置、空间分布,或跨时刻连续变化的轨迹状态。对于声源定位与跟踪任务而言,二者在目标上相互关联且在处理流程上具有递进关系。声源定位主要关注单帧或短时窗口内声源方向、位置等空间信息的瞬时估计,为后续时序分析提供基础观测;声源跟踪则通常建立在定位结果之上,进一步结合目标运动的时间连续性,对跨时刻的状态变化进行关联、平滑与更新,从而获得更加稳定和连续的轨迹估计结果。总体而言,基于深度学习的方法在一定程度上将传统声源定位与跟踪中相对分离的特征提取、状态估计与决策过程整合到统一的学习框架之中,为复杂声场条件下的声源空间感知提供了新的技术路径。围绕这一总体框架,后文将进一步从输入表征、网络结构、数据集及深度学习评价指标等方面,对基于深度学习的声源定位与跟踪方法作进一步综述。

2　深度学习声源定位与跟踪的输入特征

2.1　声源定位的输入特征

深度学习声源定位的输入特征决定了模型能够利用的空间线索类型,是系统性能的关键。现有研究中,输入特征大致可归纳为:通道间特征、互相关特征、频谱图特征、Ambisonics与声强特征、原始波形特征以及其他补充性特征。不同特征从不同角度描述声场结构,常被组合使用以提升定位精度与鲁棒性。

(1)通道间特征。通过不同麦克风之间的相对差异表达声源方向,是最基础的空间线索。相对传递函数(Relative transfer function, RTF)利用麦克风对的频域比值编码传播路径,其相位与时延相关,幅度反映能量差异。双耳特征源自人类听觉机制[39],能够有效表征方位与高程信息,尤其适用于双通道系统。通道间特征物理意义明确,但在多源或强混响条件下估计难度增加。

(2) 互相关特征。通过度量通道间相似性捕获传播延迟是传统TDOA方法的延续。GCC-PHAT是最常用的形式[6],通过相位加权增强时延峰值,适合与深度模型结合以提升噪声环境下的稳定性。交叉功率谱(Cross-power spectrum, CPS)在频域中同时保留幅度与相位信息,适合与时频建模结构结合。传统空间谱方法(如SRP-PHAT、MUSIC)也可作为输入,将空间能量分布直接提供给网络,融合了物理先验。

(3) 频谱图特征。直接使用多通道短时傅里叶变换(Short-time Fourier transform, STFT)表示,让网络从原始时频结构中学习空间差异,是深度学习声源定位中最常见的特征形式[40]。幅度谱图易处理但空间信息有限;相位谱图包含丰富方向线索;复谱图通过实部与虚部避免相位包裹问题[41];梅尔(Mel)频谱或Bark频谱等基于感知尺度的谱图能够有效降低特征维度,因而更适用于轻量化模型。尽管频谱图特征具有较强的表征能力,但其计算量通常也相对较大。

(4) Ambisonics。与声强特征提供阵列无关的空间表示,适用于不同麦克风配置。Ambisonics通过球谐函数分解声场[42],高阶形式具有更高方向分辨率。声强特征基于声压与粒子速度的关系,其活动强度部分与声波传播方向一致,是物理意义极强的方向线索[43],对噪声与混响具有良好鲁棒性,近年来成为Ambisonics声源定位的主流输入。

(5) 原始波形特征。代表端到端声源定位的发展趋势,通过直接输入多通道波形让网络自动学习特征表示[44]。此类方法避免手工设计,但对数据规模、网络结构和训练策略要求较高,在噪声环境下稳定性仍有限,因此常与其他特征结合使用。

总体来看,声源定位的输入特征从传统信号处理逐渐演化为多尺度、物理先验与数据驱动相结合的形式。其中,通道间与互相关特征提供明确空间线索,频谱图特征支持网络自动学习,Ambisonics与声强特征提供阵列无关表示,而原始波形则推动了端到端建模。未来趋势包括更强的阵列无关性、更轻量化的空间特征以及多模态融合。

2.2 声源跟踪的输入特征

与定位相比,声源跟踪不仅关注声源在某一时刻的位置,还需要刻画其随时间的运动轨迹。因此,跟踪任务在定位特征的基础上,引入时间维度建模,使输入特征具有显式或隐式的时序结构[22]。

一类常见方式是对上述空间特征进行多帧时间堆叠,例如构建连续帧的直达路径相位差(Inter-channel phase difference, IPD)、通道间电平差(Inter-channel level difference, ILD)、通道间时延差(Inter-channel time difference, ITD)、GCC或STFT特征序列,以学习方向随时间的平滑变化规律[45]。

另一类方法则利用空间谱时间序列,如连续SRP-PHAT或MUSIC空间谱图,通过学习谱峰随时间的移动轨迹实现多目标跟踪。在联合声源事件检测与定位(Sound event localization and detection, SELD)框架中,多通道Mel特征或Ambisonics强度向量通常以时间序列形式输入模型,同时预测声源存在概率与方向轨迹[46]。

因此,跟踪任务的输入特征本质上是在空间特征基础上的时序扩展,其核心在于“时空联合建模”,强调轨迹连续性而非单帧判别能力。

3 基于深度学习的声源定位与跟踪方法研究及应用

声源定位与跟踪与深度学习的结合主要体现在利用不同类型的神经网络来学习和优化波达方向的向量。以下介绍几种主要的神经网络架构及其在声源定位与跟踪中的应用,并总结深度神经网络与真实实验结合的现状。

3.1　基于DNN的声源定位与跟踪方法及应用

DNN通常由多层全连接层堆叠而成,用于学习输入声学特征与声源空间参数之间的非线性映射。输入多为GCC-PHAT、相位谱或多通道幅度谱,输出为离散DOA类别或连续角度回归值[47]。DNN的前向传播过程可以表示为

$$\boldsymbol{h}=\mathrm{ReLU}(\boldsymbol{W}_1 x+\boldsymbol{b}_1) \tag{1}$$

$$\boldsymbol{y}=\mathrm{Softmax}(\boldsymbol{W}_2 \boldsymbol{h}+\boldsymbol{b}_2) \tag{2}$$

式中:$\boldsymbol{h}$为隐藏层激活值,$\boldsymbol{W}_1$为输入层→隐藏层权重,$\boldsymbol{b}_1$为隐藏层偏置,$\boldsymbol{y}$为输出层预测值,$\boldsymbol{W}_2$为隐藏层输出层权重,$\boldsymbol{b}_2$为输出层偏置。DNN在声源定位中主要用于:(1)将手工或传统空间特征(如GCC、ILD、ITD、相位差统计量)映射为DOA类别或角度回归输出;(2)作为定位系统的"判别器/回归器"完成方向判决。DNN在声源定位中的示意图如图2所示。

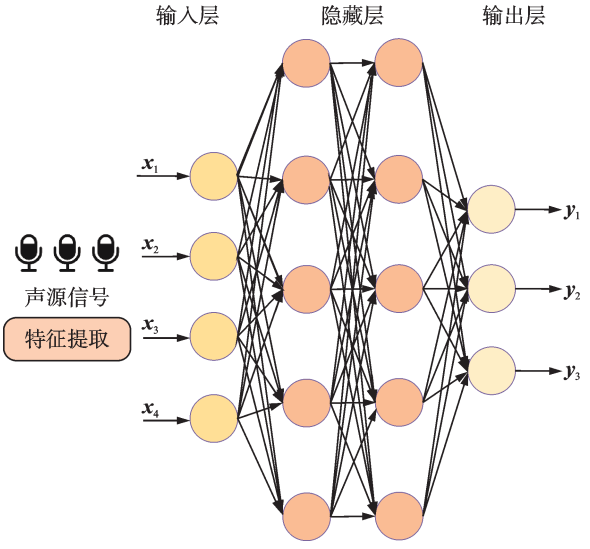

图2　基于DNN的声源定位与跟踪结构

Fig.2　Architecture of DNN-based sound source location and tracking

DNN由多层全连接结构构成,具备逼近复杂非线性映射关系的能力。在声源定位与跟踪任务中,DNN通常被用作空间特征映射算子,用以建立多通道声学特征与声源空间参数之间的对应关系,并可结合时间上下文信息实现连续帧估计。相较于依赖显式声学模型和几何假设的方法,DNN通过数据驱动方式自动学习隐含的空间判别特征,在非理想测量条件、阵列结构受限以及复杂声场环境中能够展现出更强的鲁棒性与适应性。

Chen等[48]提出了一种基于DNN的非同步测量声源定位方法,通过移动阵列构建更大的"虚拟阵列"以提升空间分辨率,但因测量时间不连续导致协方差矩阵缺失。作者设计了两种深度学习框架:基于U型网络(U-Net)的DL-NSM-U和基于残差网络(ResNet)的DL-NSM-R,并利用深度网络对非同步采样导致的不完整协方差矩阵进行补全。该方法分别对协方差矩阵的幅度与相位信息进行建模,并以同步测量条件下的完整矩阵作为训练目标,在恢复跨时段空间相关结构的同时提升了非平稳声源的定位精度,在非平稳宽带声源定位中显著优于交替方向乘子法(Alternating direction method of multipliers, ADMM)和贝叶斯分层蒙特卡罗(Bayesian hierarchial Monte Carlo, BHMC)等传统算法,为动态场景下的连续定位与声源跟踪提供了空间统计基础。。

在双耳声源定位及多声源检测任务中,DNN早期被用于直接学习空间线索与声源方位之间的映射关系。Ma等[49]采用多层全连接DNN对双耳互相关函数和声级差等高维特征进行判别建模,并结合头部转动引入多视角跟踪观测,从而有效缓解前后方位歧义并提升混响环境下多声源定位与跟踪的稳定性。同时,系统通过对连续时间块的方位后验概率进行融合,实现对声源空间位置的持续估计,为动态场景下的声源轨迹推断提供了基础。随着研究对象由单声源扩展至多声源场景,DNN进一步被用于高分辨率空间建模与未知源数条件下的联合检测与定位,在机器人真实录音环境中取得了较高的检测与定位精度,体现出其在多声源连续定位与潜在跟踪任务中的应用潜力。

Wang等[50]提出了一种基于层级多尺度骨干网络(Res2Net)与动态卷积的声源定位框架的多尺度

特征注意力网络(Multi-feature attention network, MFAnet)网络,在DNN框架中融合Res2Net结构、动态卷积以及空间与通道注意力机制,并通过多输出回归直接估计声源的三维空间坐标。与此同时,DNN也被用于与传统阵列信号处理方法相结合,例如在球谐域声源定位中通过学习嵌入特征并指导波束形成权重估计,或在基于物理信息神经网络的分布式声学传感系统系统中结合几何求解器完成位置反演,这类方法在高混响与低信噪比条件下仍能保持较高的定位精度,体现了数据驱动模型与物理建模协同的优势。

针对深度学习声源定位对大规模标注数据依赖较强的问题,Liu等[51]提出了一种基于距离度量学习的声源定位方法。该方法通过核化距离度量学习对RTF特征进行非线性映射,并引入类间分离与复杂度约束正则项,以增强DOA分类判别性。在仅使用1 482条标注样本的条件下,该方法在多种非匹配噪声与混响环境中取得了约82.17%的平均定位准确率,相比CNN与半监督深度模型在低资源场景下表现出更强的稳定性。

Yang等[52]提出基于深度神经网络的SRP-DNN框架,通过因果卷积递归网络学习多麦克风对的时变直达路径相位差序列,实现对多运动声源的连续空间估计。该方法利用循环结构建模声源轨迹的时间连续性,并结合迭代主源检测与去除策略实现逐帧分离与动态跟踪。在LOCATA数据集上,单运动声源场景下漏检率仅为0.1%,平均方位误差为2.5°;双运动声源场景中漏检率由传统SRP-PHAT的25.5%降至7.4%,平均方位误差保持在2.8°。结果表明,DNN在复杂噪声与混响条件下能够实现稳定的动态定位与轨迹估计。

基于深度神经网络的声源定位与跟踪方法在多声源、强混响及复杂噪声环境下已展现出显著优势,但其发展仍面临若干关键问题。首先,需进一步提升DNN的跨场景泛化能力,减少对特定阵列结构、声源数量及声场统计假设的依赖,这在现有多声源检测方法中尤为迫切。其次,如何在保证定位精度的同时降低训练与推理开销,是推动DNN实时方法走向实时应用和嵌入式部署的关键。最后,DNN从单帧或短时决策扩展至更长时间尺度的连续跟踪建模,并与后端滤波或运动模型协同设计,仍是提升声源定位与跟踪系统鲁棒性和实用性的一个重要方向。

3.2 基于CNN的声源定位与跟踪方法及应用

CNN通过卷积核在时频维度上提取局部空间相关特征,常用于建模多通道信号中的相位差、幅度差及其邻域关系[53]。CNN在声源定位示意图如图3所示。

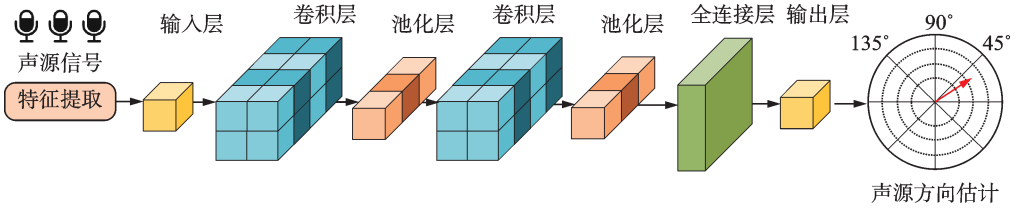

图3 基于CNN的声源定位与跟踪结构

Fig.3 Architecture of CNN-based sound source location and tracking

CNN形式可以表示为

$$\boldsymbol{H}_{l+1}=f(\boldsymbol{W}_l\cdot\boldsymbol{H}_l+\boldsymbol{b}_l) \tag{3}$$

式中:$\boldsymbol{H}_l$为第l层的特征,$\boldsymbol{W}_l$为卷积核,$\boldsymbol{b}_l$为输出层偏置。CNN在声源定位的作用有:(1)从多通道时频特征中自动提取与DOA强相关的局部相位差与幅度差分布;(2)在短时窗口内输出DOA概率或角度估计,实现静态/弱动态定位。CNN通过局部感受野与参数共享机制,能够从多通道或双耳信号中高效

提取稳定的空间判别特征,是深度学习声源定位与跟踪中最早被系统应用,也是使用最广泛的模型之一。

Chen等[54]提出了一种端到端深度学习框架,实现多声源分离与定位。该方法通过联合训练分离网络与TDOA估计网络,并引入混合信号重构结构与判别器,有效提升分离质量与定位精度。实验采用基于受限资源卷积神经网络的子带域分离(Sub-band domain resource-constrained mixture recovery network, SuDoRM-RF)与双路径变压器网络(Dual-path transformer network, DPTNet)作为分离器,结果显示分离性能提升1.3～2.7 dB,DOA平均误差降低至1.67°,在仅使用四麦克风的条件下即可达到或超越部分八麦克风方法。

在双耳极简阵列场景下,Geva等[55]提出了一种基于CNN的双耳声源定位与跟踪混合模型,该方法同时利用时域波形的时间与相位信息以及频域谱特征,结合耳廓相关的头相关传递函数(Head-related transfer function, HRTF),实现全空间高精度定位。模型由波形单元、频谱单元及融合层组成,采用端到端训练方式,并在损失函数中结合欧氏距离与角度误差以优化方向估计。由于模型可对连续时间片输入进行方向回归预测,因此能够通过逐帧位置更新形成声源运动轨迹,实现双耳条件下的动态声源跟踪。实验基于KU100假人头与24扬声器采集的真实数据,平均角度误差仅0.24°,空间误差0.01 m,显著优于基准模型(19.07°,1.08 m),在小型阵列条件下为高精度定位与跟踪提供了有效方案。

针对多声源重叠和强混响环境,部分研究通过结构改进增强CNN的表达能力,Wang等[50]提出了MFANet,该方法通过在Res2Net残差模块中引入动态卷积,使卷积核权重能够根据输入自适应调整,从而增强频率与时间域特征建模能力。同时,网络在特征提取阶段嵌入空间与通道双注意力机制,可有效提升模型对关键特征的聚焦能力。最终,利用双向门控循环单元(Bidirectional gated recurrent unit, BiGRU)与全连接层进行多输出回归,实现声源三维坐标估计。实验在ANSYN、RESYN与REAL数据集上验证了其性能,结果表明,MFANet在复杂声学环境和多声源重叠场景下均优于现有方法,尤其在真实数据集上表现出更高的定位精度与检测率。此外,该模型参数量仅1.76×10^{6},推理速度达58 fps,具备实时性与轻量化优势。研究展示了改进卷积结构与注意力机制在声源定位与跟踪中的有效性与应用潜力。

文献[56]提出了一种基于轻量化CNN的双耳多声源定位与跟踪方法,该方法以轻量化倒置残差结构网络(Inverted residual mobileNetV2, MobileNetV2)为主干,对ITD、ILD等双耳空间特征进行卷积建模,并将多声源定位与计数统一为帧级多标签分类任务。该方法实验结果表明,在合成数据集上定位跟踪帧级延迟约38.8 ms,具备良好的实时性。但在真实混响和多声源条件下,定位精度仍有待进一步提升。

Bozkurtlar等[57]提出了一种改进型CNN声源定位与跟踪方法,在ResNet框架下引入Von Mises-Bernoulli激活函数以增强对相位周期性特征的建模能力。该方法以多通道相位谱为输入,通过卷积与残差结构学习声源方向相关的空间特征,并将DOA估计建模为角度分类任务。模型以固定时间块进行实时预测,连续输出逐帧DOA序列,从而通过时间序列上的方向变化实现对声源运动轨迹的隐式跟踪。在有噪条件下,其定位误差较标准ResNet降低约50%,体现了在动态声场中连续定位与跟踪的有效性。

在文献[58]中,Tang等提出的基于CNN的声源定位方法直接以多通道时间域信号作为CNN网络输入,通过多层卷积结构联合建模通道间的时延与幅值差异,实现了端到端的声源位置预测。实验结果表明,在单声源条件下定位准确率可达到95%以上,在特定信号类型下最高可达100%,同时具有较低的旁瓣干扰。但该方法对训练数据规模与阵列配置较为敏感,在多声源及复杂声场中的泛化能力仍有提升空间。

卷积循环神经网络(Convolutional RNN, CRNN)通过在CNN提取空间特征的基础上引入循环结构对时间序列进行建模,能够显式刻画声源随时间变化的运动轨迹与声事件持续性,在动态声场和多声源场景中对声源进行定位与跟踪具有明显优势。Adavanne等[22]首次系统性地将CRNN引入SELD任务,该方法通过CNN学习多通道时频特征中的空间判别表示,并利用RNN对声事件在时间维度上的连续演化进行建模,从而实现在逐帧输出中同时估计多个声源的方向信息以及对动态声源的时序跟踪。该方法采用空间噪声信号而非语音作为训练输入,使模型更专注于阵列几何与空间结构特征,从而减弱语音内容差异的影响。在多声源仿真场景中,该方法在相邻声源角度间隔较小时仍保持稳定定位与跟踪性能,平均定位误差较传统空间谱方法降低15%~20%。

在此基础上,Wang等[59]进一步强化了CRNN框架中的时序约束机制,通过BiGRU对连续帧声事件进行建模,同时进行声事件检测(Sound event detection, SED)与DOA估计,并在DOA分支中引入基于历史预测的动量更新策略,利用声事件的时间一致性来平滑定位结果,降低误差;同时在损失函数中加入L_2正则化项,缓解小数据集上的过拟合问题。实验结果表明,该方法在DCASE 2022低资源真实数据集上显著降低了跟踪误差并提升了定位召回率。

在CNN中,U-Net采用编码-解码结构与跳跃连接机制,能够在融合多尺度上下文信息的同时保留空间细节,因而被用于声源空间分布建模与多任务联合优化。在文献[60]中,U-Net对多通道时频特征进行端到端建模,将声源定位、声事件检测与噪声抑制统一到同一框架中,通过跳跃连接保持细粒度方向线索,使定位与检测任务在同一网络中协同学习。该模型在时间维度上对连续帧进行联合预测,输出逐帧DOA与事件活动状态,从而在声源移动或事件持续变化时形成稳定的方向轨迹,实现定位与跟踪的一体化建模。在DCASE数据集上,该方法取得约8°的定位误差和0.65的F_1-score,整体性能优于传统CNN方法,表明多尺度特征融合与联合优化机制有助于提升动态场景下声源定位与跟踪的稳定性与鲁棒性。

而在文献[61]中,基于U-Net的密集连接卷积网络在单声源场景下被用于从低分辨率波束形成声源图中恢复高分辨率空间分布,并结合动态卷积与空间、通道注意力机制,对卷积特征进行自适应调节,降低计算复杂度。实验结果表明,该方法在REAL数据集上显著降低了定位误差,检测准确率达到93%以上。

展望未来,基于CNN的声源定位与跟踪研究仍具有广阔的发展空间。一方面,如何在保持轻量化结构与实时性能的同时,进一步提升CNN在多声源重叠、强混响及低信噪比条件下的定位稳定性,将是重要研究方向。现有研究表明,多尺度卷积、致密连接及相位感知卷积等结构改进能够在一定程度上缓解复杂声场下的性能退化,但其泛化能力仍依赖训练数据分布,亟需结合更具物理意义的声学约束或自监督学习策略加以改进。另一方面,面向连续声源跟踪任务,单纯的帧级CNN预测难以充分刻画声源的时序演化过程,未来可通过在CNN框架内引入轻量级时序建模或跨帧特征聚合机制,实现定位与跟踪的协同优化。此外,随着嵌入式与边缘计算需求的增长,面向特定阵列结构的结构剪枝、模型压缩及阵列无关建模方法,将进一步推动CNN在实际声源定位与跟踪系统中的落地应用。

3.3 基于RNN和LSTM的声源定位与跟踪方法及应用

RNN用递归状态h_t汇聚历史信息,适合处理“帧序列”形式的声源运动与跟踪[62]。计算过程表示为

$$\boldsymbol{h}_t = f(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1} + \boldsymbol{b}) \tag{4}$$

$$\boldsymbol{y}_t = g(\boldsymbol{W}_y \boldsymbol{h}_t + \boldsymbol{b}_y) \tag{5}$$

式中:$g(\cdot)$为输出映射函数,$\boldsymbol{W}_x$、$\boldsymbol{W}_h$和$\boldsymbol{W}_y$为权重矩阵,$\boldsymbol{b}$为偏置项,$\boldsymbol{y}$为输出。在声源定位中,RNN的

作用有:(1)建模声源方向随时间变化的动态演化过程;(2)利用历史观测信息对声源位置进行连续时间估计;(3)实现声源定位结果在时间维度上的平滑与跟踪。RNN在声源定位与跟踪示意图如图4所示。

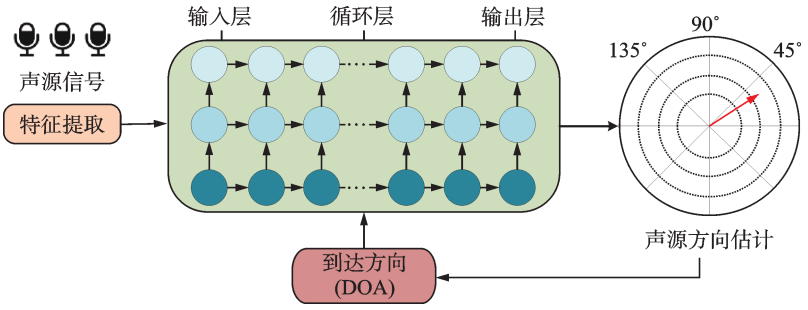

图4　基于RNN的声源定位与跟踪结构

Fig.4　Architecture of RNN-based sound source location and tracking

LSTM的出现是为了解决RNN的长期记忆遗忘问题,其核心计算公式为

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{6}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{7}$$

$$\tilde{\boldsymbol{C}}_t=\tanh(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C) \tag{8}$$

式中:σ为Sigmoid激活函数,$\boldsymbol{W}_f$为权重矩阵,$\boldsymbol{h}_{t-1}$表示前一时刻的隐藏状态,$\boldsymbol{x}_t$为当前输入,$\boldsymbol{b}_f$为偏置项,$\boldsymbol{i}_t$控制新信息的比例,$\tilde{\boldsymbol{C}}_t$生成候选记忆单元[63]。LSTM在声源定位中主要用于:(1)捕获声源时序演化;(2)预测声源运动轨迹;(3)实现连续方位跟踪。LSTM网络在声源定位与跟踪的结构图如图5所示。

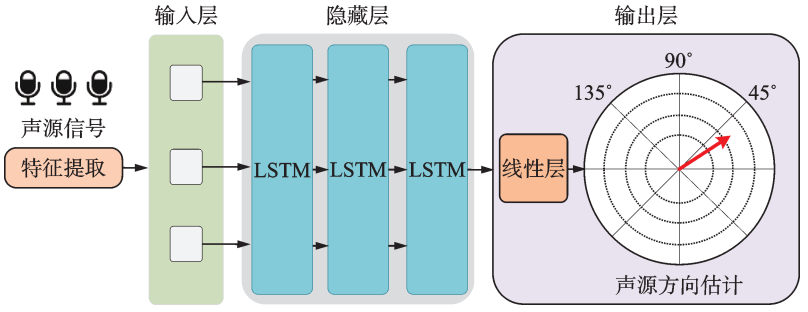

图5　基于LSTM的声源定位与跟踪结构

Fig.5　Architecture of LSTM-based sound source location and tracking

RNN过显式引入时间递归结构,使模型在当前时刻的定位与跟踪估计中能够利用历史帧信息,从而在时间维度上对声源位置变化进行连续建模。相较于基于单帧特征的静态方法,这类结构更有利于抑制瞬时估计波动,提高跟踪定位结果在时间轴上的一致性,因此在声源缓慢移动或短时动态变化的场景中具有天然优势。

已有研究系统分析了时序建模对定位稳定性的影响。围绕移动声源的方向估计问题,Rusrus等[64]系统研究了不同深度学习架构在移动声源DOA估计中的表现,通过构建模拟房间环境,生成含噪声与混响的多通道移动声源数据,并采用RNN对连续时间帧的STFT幅度比与相位差特征进行序列建模,

从而显式利用声源运动过程中的时间相关性,实现对移动声源方向的连续估计与轨迹跟踪。研究同时分析窗口大小、序列长度等参数对性能的影响,并比较了前馈神经网络、循环神经网络以及时间卷积网络,揭示了特征选择与架构差异对移动声源定位精度的关键作用,为深度学习在复杂动态场景下的声源定位与跟踪提供了重要参考。

类似地,Yang等[52]设计了一个因果CRNN,采用单向门控循环单元(Gated recurrent unit, GRU)对IPD序列进行因果时序建模,显式利用声源连续运动所带来的时间平滑特性。通过多帧信息的累积与动态建模,网络能够抑制混响与噪声引起的瞬时失真,使相位差估计在时间维度上保持一致性,从而提升移动声源的连续定位稳定性。为避免多目标输出维度不确定与分配歧义,作者将源活动编码为权重并将直达路径IPD编码为加权和形式,实现对多个移动声源的统一建模。预测得到的IPD用于构建改进的空间谱,并结合迭代检测与主导声源剔除策略分离峰值,实现多声源的逐帧检测与连续跟踪。实验在仿真数据与LOCATA真实数据集上验证,该方法在噪声和混响条件下显著优于SRP-PHAT及其他基线方法,体现了RNN在建模动态声源轨迹方面的优势。

在SELD任务中,RNN更常作为卷积特征之后的时序建模模块,用于联合刻画声事件的时间演化与空间轨迹特性。如Wang等[59]提出了一种面向低资源真实数据的SELD损失函数设计方法,该模型采用双分支结构,同时进行事件检测与方向估计,并引入辅助分类网络提供全局事件信息以增强鲁棒性。在DOA分支中,作者提出了动量更新策略,利用声事件的时间一致性来平滑定位结果,从而形成稳定的方向轨迹,实现对活动声源的连续跟踪;同时在损失函数中加入L_2正则化项,缓解小数据集上的过拟合问题。实验在DCASE 2022 Task 3数据集上验证了该方法的有效性,结果显示在定位误差与检测精度方面均显著优于基线模型,并且在不同网络架构下均能带来一致改进。然而,该方法仍主要依赖小规模数据集验证,在更复杂、多源的真实场景中的泛化能力尚需进一步探索。

进一步地,Sato等[65]同样在CRNN-SELD框架下引入RNN时序建模,并结合声功率级(Acoustic power level, PWL)先验与声衰减物理模型,实现声源距离的连续估计,在真实录音数据上有效降低了距离估计退化程度。系统利用典型类别PWL作为先验,并通过自适应声压级网络(Adaptive sound pressure level network, ASPLNet)与混响估计网络(Reverberation estimation network, RevNet)分别估计时间变化的功率级与房间声学参数,从而在数据驱动与物理建模之间建立联系。实验在合成数据与STARSS23真实数据集上验证了方法的有效性,在真实场景中显著降低距离估计退化程度,实现更稳定的三维定位与连续跟踪性能。

由于标准RNN在长序列建模中易受梯度消失问题影响,部分工作进一步采用LSTM增强对长期依赖关系的建模能力。如文献[66]将小波散射分解用于提取具有良好时频稳定性的双耳特征,并利用LSTM对特征序列进行建模,实现声源方位角与俯仰角的连续回归估计。在短时高信噪比观测条件下,该方法在训练与测试阶段均取得了接近99%的方位角定位准确率,体现了LSTM在稳定时序建模方面的优势。

基于RNN的声源定位与跟踪方法在动态声学场景中已展现出较强的时序建模能力,但结合现有文献仍存在进一步发展的空间。首先,未来研究可先围绕提升复杂运动场景下的建模能力展开,通过引入多尺度时间建模或与并行时序结构融合,缓解RNN在高速运动和长时序条件下性能受限的问题。其次,增强跨场景泛化能力将成为关键方向,现有研究表明RNN与CRNN在合成与真实数据间存在显著性能落差,未来可结合物理先验、自监督学习或跨域适配策略,以提高模型对真实声场分布变化的适应性。此外,降低计算开销与提升实时性仍是工程应用的重要挑战,可通过轻量化循环单元、序列裁剪或与非递归时序模型协同设计,实现性能与效率的平衡。最后,RNN有望在多模态与多任务联合建模中持续发挥作用,通过与视觉、语义信息的深度融合,进一步提升声源轨迹估计在复杂真实环境中的稳定性与可靠性。

3.4　基于Transformer的声源定位与跟踪方法及应用

标准Transformer由编码器和解码器两部分构成,二者均由多层结构相同的网络单元堆叠而成。每个网络单元通过注意力计算对输入序列进行加权重组,并结合前馈映射对特征进行非线性变换。为弥补模型对序列顺序不敏感的问题,Transformer通过引入位置相关表示,使模型能够感知元素之间的相对或绝对位置关系[67]。由于计算过程不依赖时间递推,Transformer可在训练阶段实现高度并行化,在长序列建模任务中展现出较高的效率和表达能力。Transformer在声源定位与跟踪的示意图如图6所示。

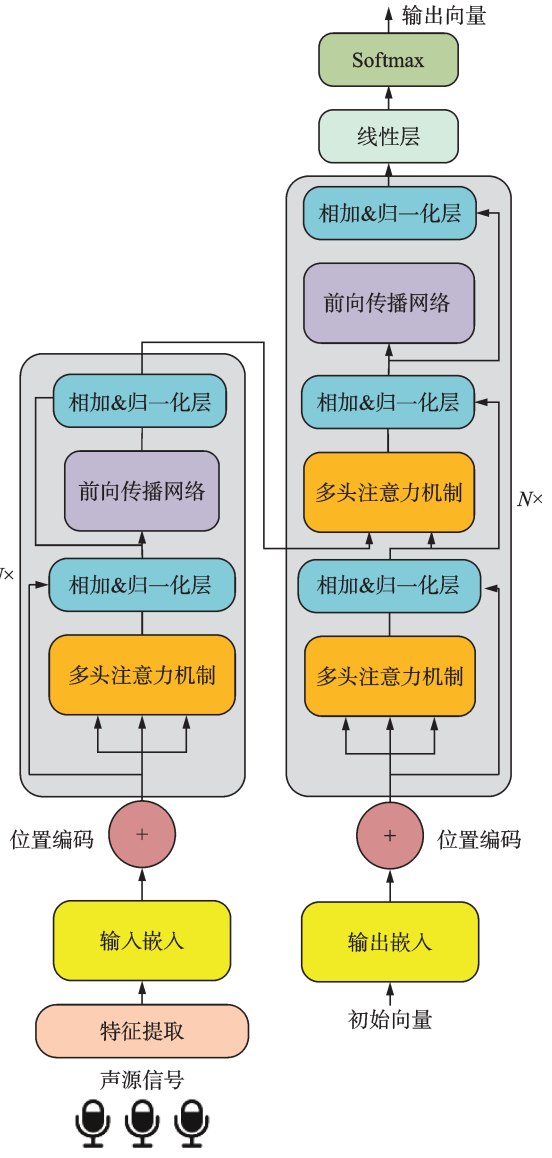

图6　基于Transformer的声源定位与跟踪结构

Fig.6　Architecture of Transformer-based sound source location and tracking

基于Transformer的声源定位方法的核心优势在于能够在全时间范围内联合建模多帧声学观测信息,通过注意力机制自适应地聚焦于与声源方向相关的关键时间位置,从而提升方向估计的整体一致性。相比依赖局部或短时建模的传统方法,Transformer更擅长刻画声源随时间变化的全局特性,对非平稳信号和动态声源具有更强的表征能力。同时,Transformer的并行计算方式避免了逐帧递归处理,在长时间序列或高时间分辨率条件下具有更高的计算效率。此外,注意力权重能够显式反映不同时间帧对定位结果的贡献,为声源定位结果的分析与解释提供有价值的依据。这些特性使得基于Transformer的方法在多声源、强混响及复杂动态环境中具备良好的应用潜力。

在声源定位与跟踪研究中,引入Transformer的核心动机并不在于替代传统卷积结构,而是弥补其在长程时间依赖和跨通道全局关联建模方面的不足。相比基于局部感受野的CNN,Transformer通过自注意力机制显式建模序列中不同位置之间的全局关系,使模型能够同时关注时间维与通道维上的空间相关性。这一特性在多声源共存、混响显著或空间线索随时间变化的场景中尤为重要,但相应地也带来了较高的计算与存储开销。

在多模态声源定位与跟踪任务中,Transformer常被用于跨模态时空关联建模。如Qian等[26]提出基于跨模态注意力融合模块(Cross-modal attention fusion, CMAF)的音视频网络,通过对音频与视觉序列分别进行自注意力建模以捕获各模态的时间依赖,并利用跨模态注意力实现音频DOA特征与视觉人脸位置信息的对齐与互补,从而提升复杂场景下的空间感知鲁棒性。该模型在连续帧序列上进行联合建模,能够利用声源运动的时间相关性实现平滑的方向估计,为说话人轨迹跟踪提供支持。实验结果表明,该方法在真实机器人场景中有效降低了平

均角度误差,将定位准确率提升至80.86%,验证了Transformer在跨模态时空关联建模及声源连续定位与跟踪中的优势。

围绕SELD任务,Transformer更常以结构替换的形式融入现有框架,用于增强对长程时间依赖的建模能力。如Mao等[68]构造了一种基于ResNet与Conformer的声事件定位与检测系统(SELD-RCNet),该系统通过残差网络与Conformer模块联合建模,既能捕捉局部时频特征,又能建模长程依赖;在连续帧序列上对声源方向与事件状态进行一致性预测,为动态场景中的声源连续定位与事件级轨迹跟踪提供了关键支撑。同时设计了两个变体:SED-RCNet用于事件检测,SSL-RCNet用于定位。输入特征包括STFT幅度与相位谱、声强矢量以及对数Mel谱,分别对应定位与检测任务。为缓解训练数据不足,作者提出了Ambisonics数据增强策略,包括音频通道交换(Audio channel swapping, ACS)与时频掩蔽(Time-frequency masking, TFM),有效提升了模型鲁棒性。但其对特定场景数据的依赖仍限制了在真实开放环境中的泛化跟踪能力。

Song等[69]提出了一种双分支神经网络架构声感网络(SonicSenseNet, SSN),用于在噪声环境下同时估计TDOA和DOA,实现高精度声源定位。该方法的核心是增强型全局频率注意机制,结合全局频率加权与多头自注意力,有效捕捉时频域上下文关系。同时引入正交损失,显式解耦TDOA与DOA特征,避免相互干扰。实验结果显示其三维定位误差低于45 mm,该研究为动态声源定位与轨迹跟踪提供了有效框架。

Zhang等[70]设计了一种动态卷积-Transformer融合神经网络(Dynamic convolution-Transformer neural network, DYCTNN)。该方法以功能波束形成(Functional beamforming, FB)图作为网络输入,通过编码-解码结构联合建模声源的空间分布信息。其中,动态卷积用于自适应提取不同声源分布条件下的局部空间特征,Transformer自注意力机制捕获空间分布的全局关联,从而增强多声源空间结构建模能力。在60通道螺旋阵列条件下,该方法在不同频率、声源数量及信噪比场景中均取得了较高的定位精度。低频条件下的定位准确率由全卷积网络(Fully convolutional network, FCN)的82%提升至96.47%,体现了引入Transformer对多声源空间建模能力的有效增强。

Kuang等[71]构建了一种融合Mamba与Transformer的双耳声源定位框架。该方法以双耳声谱为输入,通过时频patch化表示建模双耳空间线索,并采用Mamba结构替代传统自注意力模块,实现对长程时频依赖的高效建模。实验结果表明,Mamba与Transformer结合模型在噪声和混响条件下的角度误差可降低至0.89°。

综述已有工作可以发现,基于Transformer的声源定位与跟踪方法主要通过自注意力或跨模态注意力机制,加强对时序相关性、多声源干扰以及跨模态信息的建模能力,在Ambisonics-based SELD、多声事件定位及音视频联合跟踪等任务中取得了一定性能提升。然而,现有方法的优势往往建立在特定阵列形式、数据增强策略或多模态输入条件之上,其性能增益多表现为有限幅度的改进,同时伴随模型复杂度与计算开销的增加。未来相关研究可在保持Transformer时序建模优势的基础上进一步探索更高效的注意力结构设计,以降低计算成本并提升实时适用性;同时,将声学传播规律与阵列几何约束更系统地融入模型结构,有助于增强定位结果的物理一致性与稳定性。此外,针对跨阵列、跨场景条件下的泛化性能评估仍较为有限,构建统一的实验设置与评价标准,将有助于更加客观地分析Transformer在声源定位与跟踪任务中的实际价值与适用范围。

3.5 深度学习声源定位与跟踪的真实环境验证及应用

Wang等[72]构建了一种面向真实飞行无人机的深度学习辅助声源定位框架,该方法采用单通道DNN抑制无人机自噪声并估计时频掩码,再与多通道空间信息融合,构建SRP-DNN与TFS-DNN定位策略。基于AS、AVQ与DREGON等真实数据集的实验结果表明,在最低−20 dB的极低信噪比条

件下,TFS-DNN在0.5～1 s短时窗内仍保持较高检测率,显著优于传统SRP-PHAT与时频稀疏(Time-frequency sparsity, TFS)方法,验证了深度学习在真实无人机场景中的有效性与鲁棒性。

Jeong等[73]设计了一种基于可逆神经网络的声强型声源定位补偿方法,并在真实消声室实验中进行了验证。该方法采用双条件端口可逆神经网络对声强测量引入的方向性偏置进行建模与补偿。实验基于真实四麦克风阵列,结果表明,在高kd(其中k为波数,d为阵列特征尺度,二者乘积kd可表征相应的亥姆霍兹数)条件下,传统声强法误差可超过8°,而所提方法将平均DOA误差稳定压缩至1°以内,显著提升了小尺度阵列在真实测量条件下的定位精度与鲁棒性。

Jeon等[74]针对非视距道路场景中车辆难以被传统视觉与雷达感知的问题,提出了一种融合深度学习与声学建模的声源定位与跟踪框架。该方法基于多通道麦克风阵列采集的真实车辆声音,结合SRP-PHAT特征与学习型似然建模,在粒子滤波框架下构建状态递推模型,通过时间上的预测与更新实现对车辆位置的连续估计,从而完成动态轨迹跟踪。作者在自建的ARIL实车数据集及OVAD真实道路数据集上进行验证,在典型T字路口场景中实现了米级定位精度,相比传统声学方法显著降低了误差,验证了深度学习辅助声源定位与跟踪在真实交通环境中的有效性。

Lee等[75]提出了一种面向球形麦克风阵列的深度学习声源定位方法,并在真实阵列系统上进行了实验验证。该方法利用深度神经网络对球谐域相关特征进行建模,在保持较低计算复杂度的同时提升角分辨率与定位速度。真实实验结果表明,相比传统SRP/MUSIC类方法,该方法在多声源条件下显著降低了DOA估计误差,并实现了更快的定位响应。

Ma等[49]提出了一种结合深度神经网络与头部运动信息的双耳声源定位与跟踪方法。该方法基于真实测量的头相关脉冲响应(Head-related impulse response, HRIR)和双耳房间脉冲响应(Binaural room impulse response, BRIR)数据,在多个真实房间混响环境中(混响时间最高0.89 s)对多声源场景进行验证。实验结果表明,在一至三声源条件下,该方法在全方位角范围内可实现96%的定位准确率,并显著降低前后混淆,体现了深度学习在真实双耳混响环境中声源定位的鲁棒性。

Geva等[55]将时域波形与频域谱特征融合,设计深度学习双耳声源定位方法。该方法基于IRCAM录音棚中使用KU100假人头采集的真实HRIR数据进行验证,覆盖24个三维空间声源位置。实验结果表明,其平均角度误差仅为0.24°,显著优于对比模型的19.07°,在全空间范围内实现了高精度定位,体现了深度学习方法在真实双耳声场实验中的有效性。

赵东阳等[76]提出了一种面向分布式声波传感系统的物理信息神经网络声源定位与跟踪框架,并在真实Φ-OTDR系统上进行了实验验证。该方法将TDOA物理模型以解耦方式嵌入深度学习训练过程,在室外开放环境下采集6 570条真实数据进行评估。实验结果表明,其平均定位误差为4.19 cm,较纯数据驱动模型提升37.7%,在10～11 dB噪声条件下误差最高降低54.3%,体现了物理信息约束在复杂环境下对声源定位与跟踪稳定性的增强作用。

文献[77]设计了一种基于对比学习的多模态声事件定位与跟踪方法,并在真实音视频场景录制数据上进行了实验验证。该方法通过构建音频与视觉模态之间的对比表示学习机制,实现跨模态特征的一致性建模,从而提升复杂环境下的定位与跟踪鲁棒性。实验基于真实室内与户外多声源音视频数据集开展,在多声源重叠与遮挡条件下,相比仅使用音频模态的方法,事件定位F-score提升5%～10%,DOA角度误差明显降低,验证了多模态对比学习在真实复杂场景中的有效性。

文献[78]针对声源位于结构内部、传感器布置在结构表面的实际应用场景,设计了一种基于对抗式域适配的深度学习声源定位方法,并在真实结构体实验环境中进行了验证。该方法利用辅助分类器循环生成对抗网络(Auxiliary classifier cycle-consistent GAN, Ac-CycleGAN)在仿真数据与真实测量数据之间建立无配对映射关系,使模型能够在有限真实数据条件下完成结构内部声源定位。实验基于真

实结构体上的加速度传感器实测振动信号开展，结果表明，相比未进行域适配的模型平均误差下降约30%。

综上，真实实验研究表明，深度学习声源定位与跟踪方法在复杂实际声场中具有显著优势。与仿真环境相比，真实场景中普遍存在设备自噪声、混响、多径传播及传感器误差等问题，而深度学习模型通过数据驱动学习能够有效建模这些非理想因素，从而提升低信噪比及非视距条件下的定位稳定性。同时，多项研究通过消声室、真实道路、无人机飞行、双耳录音棚及结构体测试等多种真实实验平台验证了方法的可行性，表明模型在不同阵列形式与应用场景下均具备良好的鲁棒性。此外，将物理模型补偿、多模态信息融合及域适配策略引入真实实验流程，有助于进一步缩小仿真与实际应用之间的性能差距。总体而言，真实实验结果一致证明，深度学习方法能够在复杂真实环境中实现更稳定、更高精度的声源定位与跟踪，但其跨设备、跨场景的一致性表现仍需进一步通过大规模真实数据加以验证。

为了更直观地展示不同深度学习网络架构在声源定位与跟踪应用中的性能表现，表1汇总了主要研究工作的关键技术特征以及优缺点。

表1 基于深度学习的声源定位与跟踪方法性能对比

Table 1 Performance comparison of deep learning-based sound source location and tracking methods

网络模型	主要实例	技术特点	优势	不足
BMSSLNet	Wang等[56]提出的基于CNN的声源定位与跟踪方法	(1)双耳空间特征驱动卷积建模 (2)帧级多声源定位与计数联合 (3)轻量CNN支持低延迟推理	(1)帧级多声源联合定位与计数能力 (2)高效双耳空间特征利用 (3)低延迟实时推理性能	(1)对复杂声场适应性有限 (2)多声源场景精度有所下降 (3)时序上下文利用较少
DenseNet-style CNN	Zhang等[61]提出的稠密连接CNN方法	(1)声学成像结果驱动 (2)致密连接融合多尺度特征 (3)端到端预测高分辨率声源图	(1)高分辨率声源空间图重建能力 (2)多尺度特征融合 (3)显著优于传统声学成像的定位精度	(1)场景假设受限 (2)计算复杂度仍有优化空间 (3)对前端成像依赖较强
CRNN+RNN based SELD	Sato等[65]提出的RNN与物理信息先验结合的方法	(1)RNN建模事件时序 (2)引入声功率物理先验 (3)联合定位与距离估计	(1)距离估计更稳定 (2)泛化性能明显提升 (3)物理一致性更强	(1)依赖先验准确性 (2)结构复杂度较高
MP-DNN	He等[18]提出的DNN多峰输出建模的方法	(1)多峰输出建模多声源 (2)检测与定位联合学习	(1)支持未知声源数 (2)真实机器人实验验证	(1)强依赖特定阵列 (2)跨场景泛化有限
CNN+Conformer	Mao等[68]提出的融合时序建模提升多声源定位与跟踪检测性能	(1)卷积与自注意力融合 (2)长短时序联合建模 (3)Ambisonics数据增强	(1)长时序建模能力强 (2)SELD综合指标高 (3)适合复杂声事件场景	(1)依赖Ambisonics阵列 (2)模型参数规模较大 (3)实时部署成本高
SRP-DNN	Yang等[52]提出的RNN学习直达路径相位差的方法增强移动声源定位与跟踪	(1)学习直达路径相位差 (2)RNN建模空间特征演化 (3)深度模型与物理搜索融合	(1)多移动声源适应性好 (2)漏检率显著降低 (3)在线因果建模可行	(1)依赖SRP搜索模块 (2)非端到端定位流程

续表

网络模型	主要实例	技术特点	优势	不足
DYCTNN	Zhang等[70]提出的动态卷积结合Transformer的多声源时序建模定位	(1)功能波束形成声源图作为网络输入 (2)动态卷积与Transformer自注意力联合建模 (3)编码器-解码器结构实现高分辨率声源成像	(1)动态卷积增强局部空间特征表达能力 (2)自注意力机制有效建模全局空间依赖 (3)在低频与多声源场景下定位精度较高	(1)依赖功能波束形成结果作为先验输入 (2)网络结构相对复杂,训练成本较高 (3)声源时序动态建模能力相对有限
MFANet	Wang等[50]提出的多尺度残差特征结合动态卷积提升声源定位精度的方法	(1)Res2Net结合动态卷积的多尺度特征聚合 (2)频域动态卷积自适应建模时频特性 (3)空间-通道双注意力增强判别特征	(1)多尺度特征融合能力强,适应复杂声场 (2)动态卷积提升对不同声学条件的自适应性 (3)参数量小、推理速度快,具备实时部署潜力	(1)主要依赖谱图特征,空间物理先验利用有限 (2)声源时间连续性主要通过隐式建模实现 (3)对声源运动轨迹的显式建模能力不足

4　数据集和评价指标

4.1　数据集介绍

大量的音频训练数据才能使复杂的深度学习网络模型有效。为了验证所提算法对声源定位与跟踪的效果,选择合适、有效的数据集尤为重要。以下是常见的声源定位数跟踪数据集。

4.1.1　TIMIT数据集[80]

该数据集由美国国防高级研究计划局资助,并由德州仪器公司与多所研究机构共同构建,是早期具有代表性的标准英文语音语料库,包含630名说话人的朗读语音,覆盖8种主要英语方言,并提供精确的音素级时间标注。该数据集本身为单通道干净语音,但在声源定位与跟踪研究中常作为基础语音源,与房间脉冲响应和噪声模型结合生成带有空间标签的多通道数据。

4.1.2　CHiME数据集[81]

该数据集由英国谢菲尔德大学牵头,在CHiME系列国际语音挑战赛中发布,是面向真实噪声环境构建的多麦克风语音数据集,语音信号采集于多种真实日常场景中。该数据集强调真实录制条件下的多通道一致性,常用于评估声源定位与跟踪方法在强噪声和非理想声学环境中的鲁棒性。

4.1.3　LOCATA数据集[11]

该数据集由德国弗劳恩霍夫研究所等多家研究机构联合发布,是专门面向声源定位与跟踪任务构建的真实录制数据集,涵盖静态与移动声源、多种麦克风阵列结构以及不同房间声学环境。该数据集提供高精度的声源空间位置与连续运动轨迹标注,广泛用于真实场景下定位与跟踪算法的性能验证。

4.1.4　TNSSE数据集(TAU-NIGENS spatial sound events)[82]

该数据集由芬兰坦佩雷大学在DCASE SELD任务框架下发布,是针对动态声源定位与跟踪构建的合成空间音频数据集,通过将干音事件与多通道房间脉冲响应卷积生成。该数据集具有精确的方向与时间标注,适用于多声源和移动声源定位与跟踪方法的训练与测试。

4.1.5　AV16.3数据集[83]

该数据集由瑞士联邦理工学院研究人员发布,是经典的音视频多模态声源定位与跟踪数据集,包含多说话人同时存在及声源移动等复杂场景。该数据集提供连续时间的空间位置标注,常用于研究音

频与视觉信息融合的声源定位与跟踪方法。

为便于不同数据集在声源定位与跟踪任务中的适用性分析，表2从数据来源类型、声源数量、是否包含运动轨迹标注以及典型应用场景等方面对常用数据集进行统一对比。

表 2 常用声源定位与跟踪数据集对比

Table 2 Comparison of commonly used datasets for sound source localization and tracking

数据集	类型	声源	轨迹标注
TIMIT	非空间语音库(常用于合成空间数据)	单源	无
CHiME	真实录制多通道语音	单源为主	无
LOCATA	真实录制空间音频	单源＋多源	有
TNSSE	合成空间音频	多源＋动态声源	有
AV16.3	真实音视频录制	多源＋移动声源	有

4.2 评价指标

为全面评估深度学习声源定位与跟踪方法的性能，研究者通常从定位精度、时间稳定性、多声源处理能力以及实时性等多个维度构建评价指标体系。不同指标分别刻画模型在瞬时方向估计、长期轨迹跟踪、多声源场景下的匹配能力以及工程部署可行性等方面的表现。下文将围绕角度误差、轨迹误差、多声源定位误差及实时性能等常用指标，对声源定位与跟踪任务中具有代表性的评价方法进行系统介绍。

4.2.1 角度定位误差[79]

角度定位误差(Angular error, AE)用于衡量预测声源方向与真实声源方向之间的瞬时偏差，是声源定位中最基础的评价指标。在三维空间中，AE通常定义为预测DOA与真实DOA之间的角度差，其表达式为

$$\mathrm{AE}=\left|\theta_i-\hat{\theta}_i\right| \tag{9}$$

式中：θ_i表示第i个样本的真实DOA，单位通常为度(°)或弧度，$\hat{\theta}_i$为模型预测的DOA。AE于单声源与多声源定位任务中，用于刻画模型的瞬时定位精度。

4.2.2 平均角度误差/均方根角度误差

为评估模型在整个时间序列或完整声源运动轨迹上的整体定位性能，通常对角度定位误差进行统计分析。平均角度误差(Mean angular error, MAE)定义为

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|\theta_i-\hat{\theta}_i\right| \tag{10}$$

式中N表示测试样本总数。MAE能够反映模型在整体意义上的平均定位偏差。此外，为进一步强调较大定位误差对整体性能的影响，常采用角度均方根误差(Root mean square angular error, RMSAE)，其定义为

$$\mathrm{RMSAE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\theta_i-\hat{\theta}_i\right|^2} \tag{11}$$

4.2.3 定位成功率

定位成功率(Localization accuracy within an angular threshold, LA)通过统计角度误差小于给定阈值的比例来评估模型的有效定位能力，其定义为

$$\mathrm{LA}(\delta)=\frac{1}{N}\sum_{n=1}^{N}\mathbb{I}\left(\left|\theta_i-\hat{\theta}_i\right|\leqslant\delta\right) \tag{12}$$

式中:δ表示预设的角度阈值,$\mathbb{I}(\cdot)$为指示函数。

4.2.4　精确率与召回率

当声源定位与声源检测或活跃声源识别任务相结合时,常采用精确率(Precision)和召回率(Recall Rate)进行评估,其定义为

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\qquad \mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{13}$$

式中:TP、FP和FN分别表示正确检测、误报和漏检的声源数量。精确率反映模型预测结果的可靠性,而召回率衡量模型对真实声源的覆盖能力。

4.2.5　F_1-score

为综合评估模型在声源检测与定位任务中的整体性能,通常采用F_1-score指标对精确率与召回率进行联合度量,它能够在一定程度上平衡误报与漏检之间的权衡,在多声源定位与跟踪等复杂任务中被广泛采用。其定义为

$$F_1\text{-score}=\frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \tag{14}$$

5　深度学习声源定位与跟踪方法的不足与挑战

5.1　时间动态适应性与非平稳场景约束

基于深度学习的声源定位与跟踪方法在时间建模上通常依赖有限历史窗口,其适用性受到声源运动速度和序列统计特性变化的共同约束。文献[64]中基于RNN和GRU的方法在序列长度变化时表现出明显的不稳定性,递归结构在长序列非平稳输入下难以维持有效状态更新,该类模型对历史信息的依赖使其在快速变化场景中容易累积时序误差,从而限制了动态适应能力。同一研究中采用非递归结构的时序卷积网络在高速运动场景下表现出更高稳定性,但其有效性仍依赖固定时间感受野设置。在缺乏显式时间约束的情况下,文献[18]中采用帧级DNN的方法在声源运动或频繁切换时更容易产生不连续预测。由于逐帧独立建模,该类方法难以对声源轨迹变化施加连续性约束。类似地,基于Transformer的时序建模在目标快速切换或遮挡频繁的条件下难以保持稳定跟踪[65],当注意力权重随时间剧烈波动时,模型对短时干扰较为敏感。

5.2　跨数据分布与真实声场泛化约束

深度学习声源定位与跟踪方法对训练数据分布具有较强依赖性,在合成与真实声场之间普遍存在泛化约束。基于CRNN的模型在合成数据上训练后迁移至真实声场时性能明显退化,表明模型学习到的空间特征对真实声场统计变化具有较强敏感性[27]。类似现象在卷积模型中同样存在,文献[56]中BMSSLnet在真实混响条件下稳定性下降,卷积特征对混响与噪声分布变化的适应能力有限,影响了模型的通用性。动态无人机平台场景中,文献[39]中基于DNN的方法在未覆盖噪声条件下性能下降,模型对训练阶段噪声类型与强度的依赖在复杂环境中进一步放大。此外,在文献[50]中的定位与跟踪仿真实验结果显示,在真实低资源条件下采用CRNN的系统更容易出现过拟合,训练样本不足时,模型难以学习到具有泛化能力的空间表示。

5.3　多声源规模扩展与高重叠条件约束

随着声源数量增加和空间重叠程度加深,深度学习声源定位与跟踪方法在规模扩展方面面临明显约束。文献[84]指出,其基于DNN的模型在声源数量增加时定位稳定性显著下降,声源组合数量的增

长对网络判别能力形成结构性压力。而文献[56]表明,当声源重叠程度提高时,CNN对多个方向的区分能力明显减弱,局部卷积特征在高重叠条件下难以充分分离空间信息。文献[18]实验证明,采用帧级DNN的方法在多声源场景下误检和漏检现象更为突出,缺乏联合建模机制使模型难以应对声源之间的相互干扰。此外,文献[26]基于Transformer的跨模态模型在多目标频繁交互条件下性能优势减弱,注意力机制在目标密集场景中面临分配不稳定的问题。

5.4 计算复杂度与实时性约束

计算复杂度与实时性要求对深度学习声源定位与跟踪方法的工程应用构成重要约束。基于DNN的方法依赖较大规模模型和长时间训练过程[85],模型训练成本限制了其在快速部署或在线更新场景中的适用性。文献[18]进一步表明,其DNN架构对训练数据规模与计算资源具有较强依赖,该类方法在算力受限平台上实现难度较高。在卷积模型中,尽管深度复数全卷积网络在结构上提高了特征利用效率,但整体推理开销仍对实时应用构成约束[16],模型深度和连接方式增加了实现复杂度。对于注意力模型,文献[59]显示,引入Transformer风格的自注意力后模型结构更加复杂。计算负担的增加对系统实时性提出更高要求。此外,多模态方法还需额外的视频分支与跨模态计算[34],多通道输入使系统部署成本进一步上升。

综上所述,基于深度学习的声源定位与跟踪方法在时间动态变化、跨数据分布泛化、多声源规模扩展以及计算复杂度与实时性要求等方面仍面临显著挑战。现有方法的性能往往受限于训练数据分布、场景假设和网络结构选择,在应用条件发生变化时容易出现性能退化。尤其在真实声场、高动态或多声源密集场景中,上述约束更加突出,限制了方法的稳定性与通用性。如何在复杂应用环境下平衡定位精度、鲁棒性与系统开销,仍是该领域亟待解决的关键问题。

6 总结与展望

声源定位与跟踪是空间音频分析与机器听觉领域的重要基础问题,在机器人、无人系统及沉浸式音频等应用中具有重要价值。传统基于声学模型和阵列几何的方法在真实声场中易受混响、多径传播、噪声干扰及多声源重叠影响,鲁棒性受限。近年来,随着深度学习的发展,数据驱动的声源定位与跟踪方法逐渐成为研究热点。本文对不同深度学习网络模型的声源定位方法进行了分类评述,分析了优点与不足,描述了输入特征、网络结构、数据集与评价指标。现有研究表明,随着深度学习技术的发展,深度学习网络模型将会在低信噪比、强混响及多声源场景中展现出更良好的性能。对深度学习声源定位与跟踪的未来发展趋势展望如下:

(1) 声场的时序建模与智能决策能力提升

在复杂动态声场中,声源位置随时间连续变化且具有明显非平稳性,传统基于固定时序结构的深度模型在高速运动或频繁切换场景下仍易出现不稳定预测。未来可在现有深度学习框架中引入更具长程建模能力的时序模型,如基于状态空间的Mamba类结构,以弥补RNN与Transformer在长序列建模和计算效率方面的不足。同时,结合强化学习等决策机制,对声源跟踪过程进行自适应调控,有望在复杂环境下实现更稳定、连续的定位与跟踪。

(2) 泛化能力与物理一致性的协同增强

深度学习声源定位方法对训练数据分布和阵列配置依赖较强,在合成数据与真实声场之间普遍存在性能退化。未来研究可进一步将声学传播规律、阵列几何约束等物理先验融入AI模型结构或训练过程,并结合自监督与跨域学习策略,提高模型在未知环境中的泛化能力。此外,在通感一体化场景下,如何联合利用通信与感知信号,实现信息共享与协同建模,也将成为提升系统整体鲁棒性的重要方向。

(3) 多声源规模扩展与高密度场景下的协同建模

随着声源数量增加和空间重叠程度加深,现有模型在多声源定位与跟踪任务中面临精度下降和输出不稳定的问题。未来可借助图建模、多智能体强化学习等方法,显式刻画声源之间的空间与时间关联关系,提升模型在高密度声场中的扩展能力。同时,将多声源定位与声事件理解、语义分析等任务进行联合建模,有助于在复杂交互场景中实现更可靠的声源跟踪。

(4) 面向连续定位与稳定跟踪的统一表示与预测建模

未来声源定位与跟踪研究将更加关注连续时间尺度上的空间一致性与轨迹稳定性,而不仅局限于逐帧定位精度的提升。随着模型能力的增强,深度学习方法有望通过统一的空间-时间表示刻画声源运动规律,并在复杂声场中提高跟踪连续性。在此过程中,大规模预训练模型与生成式建模思想可作为辅助手段,引入声场先验与运动约束,在声源短时遮挡或状态变化时为跟踪提供预测与补全能力,从而进一步提升真实环境下声源轨迹估计的稳定性与可靠性。

参考文献:

[1] STOWELL D, WOOD M, STYLIANOU Y, et al. Bird detection in audio: A survey and a challenge[C]//Proceedings of IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP). [S.l.]: IEEE, 2016: 1-6.

[2] FOGGIA P, PETKOV N, SAGGESE A, et al. Audio surveillance of roads: A system for detecting anomalous sounds[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 17(1): 279-288.

[3] 韩雪, 李泽滔, 孙昊. 基于Kinect传感器的移动机器人声源目标跟踪系统[J]. 自动化与仪器仪表, 2015(6): 185-186.
HAN Xue, LI Zetao, SUN Hao. Sound source target tracking system for mobile robots based on Kinect sensors[J]. Automation & Instrumentation, 2015(6): 185-186.

[4] LV D, TANG W, FENG G, et al. An overview of sound source localization based condition monitoring robots[J]. ISA Transactions, 2025, 158: 537-555.

[5] QIU Y, LI B, HUANG J, et al. An analytical method for 3-D sound source localization based on a five-element microphone array[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-14.

[6] KNAPP C, CARTER G. The generalized correlation method for estimation of time delay[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 2003, 24(4): 320-327.

[7] TRAA J, WINGATE D, STEIN N D, et al. Robust source localization and enhancement with a probabilistic steered response power model[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 24(3): 493-503.

[8] ERDOGAN H, HERSHEY J R, WATANABE S, et al. Improved MVDR beamforming using single-channel mask prediction networks[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Association (Interspeech). San Francisco: ISCA, 2016: 1981-1985.

[9] TANG H. DOA estimation based on MUSIC algorithm[D]. Växjö: Linnaeus University, 2014.

[10] DANG X, ZHU H. An iteratively reweighted steered response power approach to multisource localization using a distributed microphone network[J]. The Journal of the Acoustical Society of America, 2024, 155(2): 1182-1197.

[11] EVERS C, LÖLLMANN H W, MELLMANN H, et al. The LOCATA challenge: Acoustic source localization and tracking [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1620-1643.

[12] DANG X, ZHU H. A feature-based data association method for multiple acoustic source localization in a distributed microphone array[J]. The Journal of the Acoustical Society of America, 2021, 149(1): 612-628.

[13] FUCHS J J. DOA estimation in the presence of modeling errors, the Global Matched Filter approach[C]//Proceedings of 2009 17th European Signal Processing Conference. [S.l.]: IEEE, 2009: 1963-1967.

[14] ARJOMANDI-LARI M, KARIMI M. Array auto-calibration using a generalized least-squares method[J]. AEU-International Journal of Electronics and Communications, 2019, 106: 20-31.

[15] GRONDIN F, MICHAUD F. Lightweight and optimized sound source localization and tracking methods for open and closed microphone array configurations[J]. Robotics and Autonomous Systems, 2019, 113: 63-80.

[16] GRONDIN F, GLASS J. SVD-PHAT: A fast sound source localization method[C]//Proceedings of ICASSP 2019—2019

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2019: 4140-4144.

[17] 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.

[18] HE W, MOTLICEK P, ODOBEZ J M. Deep neural networks for multiple speaker detection and localization[C]//Proceedings of 2018 IEEE International Conference on Robotics and Automation (ICRA). [S.l.]: IEEE, 2018: 74-79.

[19] NGUYEN TNT, WATCHARASUPAT KN, NGUYEN NK, et al. SALSA: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 1749-1762.

[20] VERA-DIAZ J M, PIZARRO D, MACIAS-GUARASA J. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates[J]. Sensors, 2018, 18(10): 3418.

[21] PUJOL H, BAVU E, GARCIA A. BeamLearning: An end-to-end deep learning approach for the angular localization of sound sources using raw multichannel acoustic pressure data[J]. The Journal of the Acoustical Society of America, 2021, 149(6): 4248-4263.

[22] ADAVANNE S, POLITIS A, NIKUNEN J, et al. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 13(1): 34-48.

[23] ADAVANNE S, POLITIS A, VIRTANEN T. Localization, detection and tracking of multiple moving sound sources with a convolutional recurrent neural network[J]. arXiv preprint arXiv: 1904.12769, 2019.

[24] CAO Y, KONG Q, IQBAL T, et al. Polyphonic sound event detection and localization using a two-stage strategy[J]. arXiv preprint arXiv: 1905.00268, 2019.

[25] HE C, CHENG S, BAO J, et al. Adapting single-channel pre-trained transformer models for multi-channel sound event localization and detection[C]//Proceedings of ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 1-5.

[26] QIAN X, WANG Z, WANG J, et al. Audio-visual cross-attention network for robotic speaker tracking[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 31: 550-562.

[27] FUJITA Y, BANDO Y, IMOTO K, et al. DOA-aware audio-visual self-supervised learning for sound event localization and detection[C]//Proceedings of 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). [S.l.]: IEEE, 2023: 2061-2067.

[28] YEOW J W, TAN E L, PEKSI S, et al. Environmental acoustic intelligence through sound event localization and detection: A review[J]. npj Acoustics, 2025, 1(1): 31.

[29] DIAZ-GUERRA D, MIGUEL A, BELTRAN J R. Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 29: 300-311.

[30] POLITIS A, SHIMADA K, SUDARSANAM P, et al. STARSS22: A dataset of spatial recordings of real scenes with spatiotemporal annotations of sound events[J]. arXiv preprint arXiv: 2206.01948, 2022.

[31] POLITIS A, MESAROS A, ADAVANNE S, et al. Overview and evaluation of sound event localization and detection in DCASE 2019[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 29: 684-698.

[32] DIBIASE J H. A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays[D]. Providence: Brown University, 2000.

[33] SCHMIDT R. Multiple emitter location and signal parameter estimation[J]. IEEE Transactions on Antennas and Propagation, 1986, 34(3): 276-280.

[34] CAPON J. High-resolution frequency-wavenumber spectrum analysis[J]. Proceedings of the IEEE, 2005, 57(8): 1408-1418.

[35] PAL M, ROY R, BASU J, et al. Blind source separation: A review and analysis[C]//Proceedings of International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation. [S.l.]: IEEE, 2013: 1-5.

[36] MANDEL M I, WEISS R J, ELLIS D P W. Model-based expectation-maximization source separation and localization[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 18(2): 382-394.

[37] KALMAN R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960, 82(1):35-45.

[38] RIBEIRO M I. Kalman and extended Kalman filters: Concept, derivation and properties[J]. Institute for Systems and Robotics, 2004, 43(46): 3736-3741.

[39] MAY T, VAN DE PAR S, KOHLRAUSCH A. A probabilistic model for robust localization based on a binaural auditory front-end[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(1): 1-13.

[40] CHAKRABARTY S, HABETS E A P. Broadband DOA estimation using convolutional neural networks trained with noise signals[C]//Proceedings of 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). [S.l.]: IEEE, 2017: 136-140.

[41] PEROTIN L, SERIZEL R, VINCENT E, et al. CRNN-based multiple DoA estimation using acoustic intensity features for ambisonics recordings[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(1): 22-33.

[42] POLITIS A, ADAVANNE S, VIRTANEN T. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection[J]. arXiv preprint arXiv: 2006.01919, 2020.

[43] PEROTIN L, SERIZEL R, VINCENT E, et al. CRNN-based joint azimuth and elevation localization with the ambisonics intensity vector[C]//Proceedings of 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). [S.l.]: IEEE, 2018: 241-245.

[44] TAKAHASHI N, GYGLI M, PFISTER B, et al. Deep convolutional neural networks and data augmentation for acoustic event detection[J]. arXiv preprint arXiv: 1604.07160, 2016.

[45] SHIMADA K, KOYAMA Y, TAKAHASHI S, et al. Multi-ACCDOA: Localizing and detecting overlapping sounds from the same class with auxiliary duplicating permutation invariant training[C]//Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 316-320.

[46] 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.

[47] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18 (7): 1527-1554.

[48] CHEN G, CHEN L, SUN W, et al. Deep learning aided sound source localization: A nonsynchronous measurement approach [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-15.

[49] MA N, MAY T, BROWN G J. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(12): 2444-2453.

[50] WANG R, LI Z, ZHANG B, et al. Sound source localization based on Res2Net and dynamic convolution[C]//Proceedings of 2025 4th International Symposium on Robotics, Artificial Intelligence and Information Engineering (RAIIE). [S.l.]: IEEE, 2025: 415-419.

[51] LIU M, LU Z, WANG X, et al. Sound source localization via distance metric learning with regularization[J]. Signal Processing, 2025, 227: 109721.

[52] YANG B, LIU H, LI X. SRP-DNN: Learning direct-path phase difference for multiple moving sound source localization[C]// Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 721-725.

[53] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 2002, 86(11): 2278-2324.

[54] CHEN Y, LIU B, ZHANG Z, et al. An end-to-end deep learning framework for multiple audio source separation and localization[C]//Proceedings of ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 736-740.

[55] GEVA G, WARUSFEL O, DUBNOV S, et al. Binaural sound source localization using a hybrid time and frequency domain model[C]//Proceedings of ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 8821-8825.

[56] WANG L, JIAO Z, ZHAO Q, et al. Framewise multiple sound source localization and counting using binaural spatial audio signals[C]//Proceedings of ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing

(ICASSP). [S.l.]: IEEE, 2023: 1-5.

[57] BOZKURTLAR M, YEN B, ITOYAMA K, et al. Real time sound source localization using von-Mises ResNet[C]// Proceedings of 2024 IEEE/SICE International Symposium on System Integration (SII). [S.l.]: IEEE, 2024: 466-471.

[58] TANG J, SUN X, YAN L, et al. Sound source localization method based time-domain signal feature using deep learning[J]. Applied Acoustics, 2023, 213: 109626.

[59] WANG Q, DU J, NIAN Z, et al. Loss function design for DNN-based sound event localization and detection on low-resource realistic data[C]//Proceedings of ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[60] SHIN Y, KIM Y G, CHOI C H, et al. SELD U-Net: Joint optimization of sound event localization and detection with noise reduction[J]. IEEE Access, 2023, 11: 105379-105393.

[61] ZHANG G, GENG L, CHEN X. Sound source localization method based on densely connected convolutional neural network [C]//Proceedings of 2022 5th International Conference on Information Communication and Signal Processing (ICICSP). [S.l.]: IEEE, 2022: 743-747.

[62] ELMAN J L. Finding structure in time[J]. Cognitive Science, 1990, 14(2): 179-211.

[63] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[64] RUSRUS J, SHIRMOHAMMADI S, BOUCHARD M. Characterization of moving sound sources direction-of-arrival estimation using different deep learning architectures[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 2520614.

[65] SATO N, YASUDA M, SAITO S, et al. Sound source distance estimation utilizing physics-informed prior for sound event localization and detection[C]//Proceedings of ICASSP 2025—2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2025: 1-5.

[66] MASSICOTTE P, CHAOUI H, OUAMEUR M A, et al. LSTM with scattering decomposition-based feature extraction for binaural sound source localization[C]//Proceedings of 2022 20th IEEE Interregional NEWCAS Conference (NEWCAS). [S.l.]: IEEE, 2022: 436-440.

[67] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[68] MAO Y, ZENG Y, LIU H, et al. The ICASSP 2022 L3DA22 challenge: Ensemble of ResNet-conformers with ambisonics data augmentation for sound event localization and detection[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2022: 9191-9195.

[69] SONG Y, LIU Z, ZHU Y, et al. SonicSenseNet: A dual-branch neural network with enhanced GIFA for joint TDOA-DOA sound source localization[C]//Proceedings of 2025 IEEE 15th International Conference on Signal Processing, Communications and Computing (ICSPCC). [S.l.]: IEEE, 2025: 1-6.

[70] 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.

[71] KUANG S, SHI J, VAN DER HEIJDEN K, et al. BAST-Mamba: Binaural audio spectrogram mamba transformer for binaural sound localization[J]. Neurocomputing, 2025: 130804.

[72] WANG L, CAVALLARO A. Deep-learning-assisted sound source localization from a flying drone[J]. IEEE Sensors Journal, 2022, 22(21): 20828-20838.

[73] JEONG I, PARK B, PARK K, et al. Dual-port conditional invertible neural network for sound intensity compensation in sound source localization[J]. IEEE Transactions on Instrumentation and Measurement, 2025, 74: 2504912.

[74] JEON M, CHO J K, KIM H Y, et al. Non-line-of-sight vehicle localization based on sound[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(8): 8185-8199.

[75] LEE S Y, CHANG J, LEE S. Deep learning-enabled high-resolution and fast sound source localization in spherical microphone array system[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[76] 赵东阳,万阳阳,何祖源. 基于物理信息神经网络的DAS系统二维声源定位研究[J/OL]. 光学学报: 1-17[2025-12-23], https://link.cnki.net/urlid/31.1252.O4.20251222.1118.038.

ZHAO Dongyang, WAN Yangyang, HE Zuyuan. Resesrch on two-dimensional sound source localization in DAS systems based on physics-informed neural networks[J/OL]. Acta Optica Sinica: 1-17[2025-12-23]. https://link.cnki.net/urlid/31.1252.O4.20251222.1118.038.

[77] WU S, WANG Y, JIANG Y, et al. CRATI: Contrastive representation-based multimodal sound event localization and detection[J]. Knowledge-Based Systems, 2024, 305: 112692.

[78] KITA S, PARK C S, KAJIKAWA Y. Sound source localization for source inside a structure using Ac-CycleGAN[J]. Journal of Sound and Vibration, 2024, 591: 118616.

[79] MESAROS A, HEITTOLA T, VIRTANEN T. Metrics for polyphonic sound event detection[J]. Applied Sciences, 2016, 6(6): 162.

[80] GAROFOLO J S, LAMEL L F, FISHER W M, et al. TIMIT Acoustic-phonetic continuous speech corpus[EB/OL]. (1993-05-12). https://doi.org/10.35111/zjq6-pa48.

[81] BARKER J, MARXER R, VINCENT E, et al. The third 'CHiME' speech separation and recognition challenge: Analysis and outcomes[J]. Computer Speech & Language, 2017, 46: 605-626.

[82] POLITIS A, ADAVANNE S, KRAUSE D, et al. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection[J]. arXiv preprint arXiv: 2106.06999, 2021.

[83] LATHOUD G, ODOBEZ J M, GATICA-PEREZ D. AV16.3: An audio-visual corpus for speaker localization and tracking [C]//Proceedings of International Workshop on Machine Learning for Multimodal Interaction. Berlin: Springer, 2004: 182-195.

[84] LEE S Y, CHANG J, LEE S. Deep learning-based method for multiple sound source localization with high resolution and accuracy[J]. Mechanical Systems and Signal Processing, 2021, 161: 107959.

[85] KUJAWSKI A, HEROLD G, SARRADJ E. A deep learning method for grid-free localization and quantification of sound sources[J]. The Journal of the Acoustical Society of America, 2019, 146(3): EL225-EL231.

作者简介:

陈喆(1975-),男,教授,博士生导师,研究方向:音频信号处理、图像处理和宽带无线通信,E-mail: zhechen@dlut.edu.cn。

宋登鏊(1999-),男,博士研究生,研究方向:声源定位和音频信号处理。

王一宇(2000-),女,博士研究生,研究方向:声源定位、多模态感知和音频信号处理。

殷福亮(1962-),**通信作者**,男,教授,博士生导师,研究方向:音频信号处理、图像处理和宽带无线通信,E-mail: flyin@dlut.edu.cn。

(编辑:王静)