ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 5,Sep. 2024,pp. 1085－1113
DOI:10. 16337/j. 1004-9037. 2024. 05. 004

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

分布式麦克风阵列拾音理论与应用方法综述

张　结[1]，呼　德[2]，张晓雷[3]，凌震华[1]

(1. 中国科学技术大学信息科学技术学院,合肥 230027;2. 内蒙古大学计算机学院,呼和浩特 010021;3. 西北工业大学航海学院,西安 710072)

摘　要：经过数十年的发展,麦克风阵列技术日益成熟,并广泛应用于视频会议、智能电视、移动通话和助听器等人机交互系统。然而,现实噪声或远距离交互场景中,限定阵型结构的传统麦克风阵列的拾音质量难以保证。随着无线智能终端设备的广泛使用,分布式麦克风阵列(或称无线声传感器网络)为提升复杂开放域语音交互系统的拾音质量提供了更多可能性,并在阵列组织、应用体验和声场覆盖度上更有优势。近年来,分布式麦克风阵列在很多语音交互任务上展现出良好的应用潜力,基本实现了对传统麦克风阵列语音任务的全覆盖。本文将重点总结现阶段分布式麦克风阵列的拾音理论和应用方法,包括阵列组织原理、麦克风节点效用评估,以及结合下游语音任务阐述其应用方法。最后,将简要论述分布式麦克风阵列走向实用的关键挑战与发展趋势。

关键词：分布式麦克风阵列;无线声传感器网络;麦克风效用;语音交互;拾音质量

中图分类号：TN912.3　　**文献标志码**：A

A Survey on Sound Acquisition Theories and Application Methods of Distributed Microphone Arrays

ZHANG Jie[1], HU De[2], ZHANG Xiaolei[3], LING Zhenhua[1]

(1. School of Information Science & Technology, University of Science & Technology of China, Hefei 230027, China; 2. College of Computer Science, Inner Mongolia University, Hohhot 010021, China;3. School of Marine Science & Technology, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: Over the past few decades of development, microphone array technology is becoming more mature, which has been applied to various human-machine interaction systems, e.g., video-conferencing, intelligent television, mobile telephony, hearing aids. However, in realistic noisy or distant interaction scenarios, the sound acquisition quality (SAQ) of conventional topology-constrained microphone arrays cannot be guaranteed. With the wide range of using wireless intelligent terminal devices, distributed microphone array (DMA) or so-called wireless acoustic sensor network (WASN) provides more possibilities of improving the SAQ for speech interaction systems in complex and open domains, and shows a superiority in array organization, application experience and scene coverage. Recently, DMA exhibits a good applicable potential in many speech interaction tasks, which almost cover the tasks that conventional microphone array can handle. This survey will mainly summarize some existing important sound acquisition theories and application methods of DMA, including principles of array organization, utility evaluation of

基金项目:国家自然科学基金(62101523,62201297,62361045,62176211)。
收稿日期:2024-08-05;**修订日期**:2024-09-19

microphone nodes and the application methods in combination of downstream speech tasks. Finally, we will briefly discuss some key challenges and developing trends of the road of DMA to practical usages.

Key words: distributed microphone array; wireless acoustic sensor network; microphone utility; speech interaction; sound acquisition quality

引　　言

麦克风阵列(Microphone array, MA)技术被广泛应用于语音识别、语音增强、声场监测、说话人定位、助听器、语音通信和移动机器人及视频会议等人机交互系统,其拾音质量直接关系到下游语音信息处理算法的性能。除了目标语音信号的时频域信息,麦克风阵列拾音系统还可以利用阵列拓扑结构挖掘目标声源的空间信息,因而其拾音质量和下游任务表现力相比单麦克风系统都有了明显提升[1]。

如图1所示,传统麦克风阵列是一组位于空间不同位置的全向麦克风按一定的形状规则布置形成的阵列,用于对声场声信号进行时间和空间采样。根据声源和麦克风阵列的间距(相对于Fraunhofer距离),阵列观测信号分为近场模型和远场模型。近场模型将声波看成球面波,它需要考虑麦克风阵元接收信号间的幅度差[2];远场模型则将声波看成平面波,忽略了各阵元接收信号间的幅度差,近似认为各接收信号之间仅存在延时关系,远场模型在基于麦克风阵列的波束形成技术中应用更加广泛[3]。根据麦克风阵列的拓扑结构,则可分为线性阵列、平面阵列(如十字阵列、T型阵、均匀圆阵、圆形或矩形面阵等)、立体麦克风阵列(如四面体阵、长方体阵或球型阵等)[4]。通常麦克风阵列的拓扑结构越复杂,其空间采样及声源感知能力越强[5]。

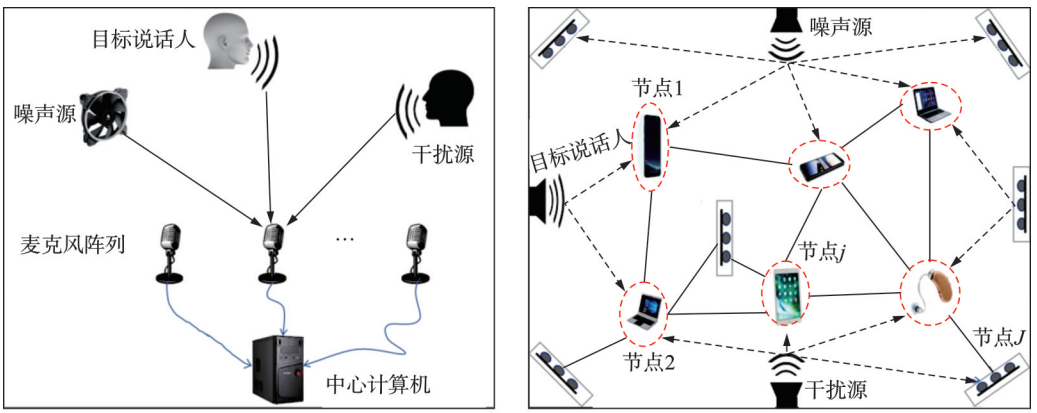

图1　传统有线麦克风阵列和分布式麦克风阵列系统示例

Fig.1　Examples of conventional wired microphone array and DMA

传统具有规则型拓扑结构的麦克风阵列存在多方面缺点。首先,在系统设计方面,由于所有麦克风需要与中心计算机进行点对点物理连接,有线阵列的重新布局(如增加或减少麦克风)不太灵活;其次,在声场感知方面,由于麦克风阵列并不能放置于空间任意位置,导致阵列的空间感知能力有限,特别是当目标说话人距离较远时;再次,在应用场景方面,虽然麦克风阵列技术已经在许多语音交互产品中得到成功应用,但是阵列尺寸严重制约了下游任务的性能,比如助听器、手持通信设备、人形机器人等系统只能搭载小型麦克风阵列[6]。因此,如何突破传统麦克风阵列的物理约束,提升其拾音质量、布控灵活性以及声场监测覆盖度已成为当前音频和语音信号处理领域的一个热点研究课题。

近年来,随着无线终端设备的普及(如智能手机、手提电脑、无线耳机及辅听器具),分布式麦克风阵列(Distributed MA, DMA)受到国内外学者的广泛关注,图1(b)展示了一种典型的基于分布式传声器阵列的语音信息处理系统。由于无线电子设备通常具有音频数据采集、无线传输和一定的移动计算能力,将声场景中的无线终端设备按照自组织(Ad-hoc)方式连接起来可构造成分布式麦克风阵列,从而替代传统有线麦克风阵列开展拾音和语音信息处理[7],对实现跨平台复杂声场景下多语音任务的并行处理具有重要潜力。分布式麦克风阵列通常也称为无线声学传感器网络(Wireless acoustic sensor network, WASN)或无线麦克风网络(Wireless microphone network)。

分布式麦克风阵列拾音系统相较于传统有线阵列具有多方面优势:首先,可拓展性更强,分布式处理可以克服对某个节点的过度依赖(如中心计算机),某个节点的短暂失效(比如断电)不会导致整个系统瘫痪,新的设备允许就近加入网络以补充计算资源;其次,阵型结构更加灵活,可中心化(Centralized)亦可去中心化(Decentralized),不再受到规则型拓扑结构的约束;另外,拾音范围更广,无线设备可能出现在声场景中任意位置,特别是当某个节点距离目标说话人或噪声源很近时(分别对目标语音提取和噪声消除非常有益),意味着有望实现全空间覆盖式均匀拾音[8],比如2022年腾讯天籁发布的inside音频解决方案[9];再次,应用场景更广,比如为受设备尺寸和功耗限制的语音处理系统(如助听器、机器人)无线共享外部传声器音频数据可以大大提升其语音感知性能[10-11];而且分布式麦克风阵列支持多种语音任务协同处理,不同节点可以重点关注不同任务(比如目标语音提取、会议转写、声源定位、噪声消除、去混响等)[12-13]。因此,分布式麦克风阵列已成为下一代声学信号采集和处理平台,在智能会议、智能家居、智慧城市、车载语音交互和助听器等场景都具有重要的应用潜力。

本文将按照"系统构建→布局优化→下游应用"的思路,重点概括现阶段分布式麦克风阵列拾音层面的组织理论和节点效用评估理论,以及相关应用方法(以语音增强/识别、声源定位、说话人识别等任务为例),分析分布式麦克风阵列在实际应用中的优势和挑战,主要内容与组织结构如图2所示。

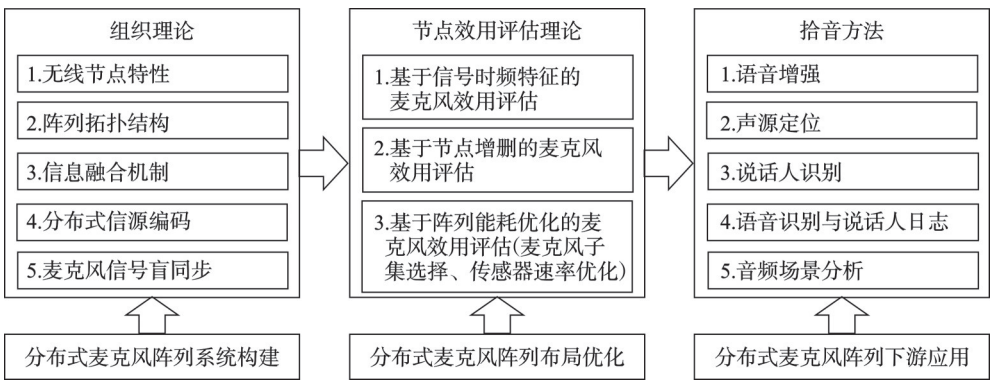

图2　本文主要内容与组织结构

Fig.2　Main contents and organization structure of this survey

1　分布式麦克风阵列组织理论

本节将从节点特性、拓扑结构、信息融合机制、信源编码和节点时钟同步几个角度系统介绍分布式麦克风阵列的组织理论,这些是构建分布式麦克风阵列拾音系统和设计下游语音交互技术的基础。

1.1　无线节点特性

随着无线终端设备的多样化发展,现实中无线节点可包含单个或多个麦克风,例如智能耳机通常使用单麦克风、手机包含双麦克风、助听器(Hearing aid, HA)可包含2～4个麦克风(即小型麦克风阵

列)、智能电视/会议平板采用线性麦克风阵列等。通常拾音麦克风数量越多,端侧拾音质量越高。另一方面,由于无线终端设备通常随机分布在声场景中,而且允许随机变化,例如用户佩戴的助听器或随身携带的手机可随着身体移动或姿态转动而位置变化,造成分布式麦克风阵列几何结构的未知性,也催生出许多麦克风节点自定位相关工作[14-17]。此外,由于麦克风节点与声源距离的随机性,导致传统远场信号模型通常并不适用于分布式麦克风阵列。从硬件角度看,无线终端通常具有有限的电池资源,无线局域网总体带宽资源也有限,这对数据通信、路由和分布式麦克风阵列算法的复杂度提出了更高的要求。因此,节点异质性(Heterogeneity)、阵列拓扑的未知性(Unknown topology)和低资源(Low-resource)是分布式麦克风阵列节点层面的三大特点。

1.2 分布式麦克风阵列拓扑结构

构建分布式麦克风阵列最直接的方式是借鉴传统麦克风阵列采用基于融合中心的集中式架构,如图3(a)所示,融合中心通常为中心计算机用于收集并处理所有节点采集和传输的音频信号。也就是说集中式网络呈星形拓扑结构,节点只负责采集和传输数据,所有复杂的信号处理任务均由融合中心完成。这种结构的优点在于网络架构简单、不涉及复杂的数据路由/中继技术,并且节点既可同步也可异步地完成数据收发任务。然而,其缺点也很明显:融合中心的负载过高且网络状态极度依赖于融合中心,中心计算机的失效会造成整个系统无法工作;数据收发能耗偏高,由于无线局域网服从IEEE 802.11通信协议标准,令麦克风节点j到融合中心的无线链路信号衰减为ρ_j(介于2~6),所有麦克风节点传输信道的噪声功率为N_0,b_j表示每个样本分配的比特数,依据高斯信道的香农定理可以推导出麦克风节点j到中心节点的每个样本发送能量为[18-20]

$$E_j = d_j^{\rho_j} N_0\left(2^{2b_j}-1\right) \tag{1}$$

即传输速率等于信道容量条件下的无损传输。式(1)表明节点发送能量与传输距离正相关,当节点距离融合中心较远时,集中式数据收发方式无疑会消耗大量能耗,这对有限电池资源储备的智能终端设备的生存期是极为不利的。

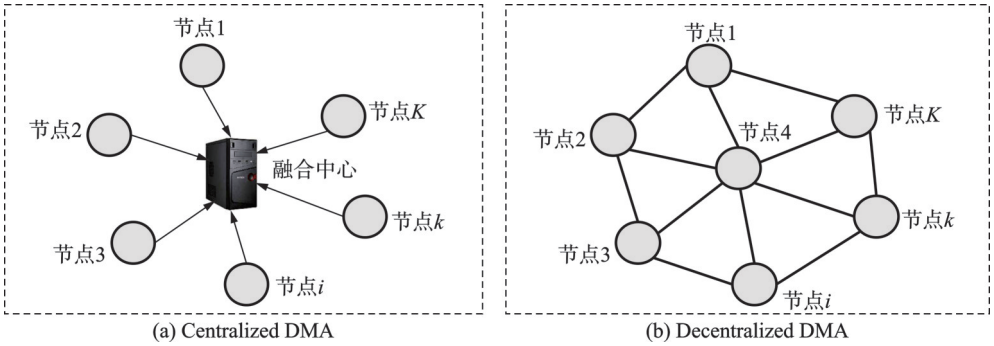

图3 集中式和非集中式分布式麦克风阵列

Fig.3 Centralized and decentralized distributed microphone array

非集中式分布式麦克风阵列移除了中心节点,允许每个节点按距离就近原则与其相邻节点通信,形成了典型的图模型$\mathcal{G}=(\mathcal{E},\mathcal{V})$,其中$\mathcal{V}$代表节点集合,$\mathcal{E}$代表边集合,当且仅当两个节点间链路存在于边集合中时才允许二者直接通信,如图3(b)所示。根据设备的安全性限制[21],例如某节点出于隐私保护的目的只允许发送数据或接收信息,还可分为有向(Directed)图和无向(Undirected)图模型。这种非集中式结构通常需要迭代式网内(In-network)数据处理,每个节点通过收集来自邻居节点的信息并结

合自身观测数据,初步进行融合处理,然后将结果无线反馈给邻居节点。显然非集中式结构的局部数据收发能耗更低、避免了对某个节点的过度依赖,网络的总体能耗和计算开销均摊到节点层面,这种方式在实际应用中也更被偏爱。然而,非集中式网络的算法设计也更复杂,因为每个节点只能获取少量的数据、节点处理顺序存在不确定性等[22]。

1.3　信息融合机制

分布式麦克风阵列通过融合和处理不同麦克风节点的观测信号,估计目标说话人语音信号,从而实现高质量拾音。其中,信息融合包括网内数据融合和分布式信号处理两个方面,本节重点讨论网内数据融合,分布式信号处理算法通常可以获得更优的信息融合效果,将在3.1节详细介绍。

中继模式:在所有节点享有充足带宽和时钟同步的理想条件下,集中式网络中节点可以直接将观测数据发送给融合中心,融合中心利用波束形成或自适应滤波技术[1,23-24]估计目标语音信号,但这种假设条件过于苛刻,现实中很难满足,但不可否认这是最优的拾音策略。一种次优的方式是采用中继(Relay)技术,如图4(a)所示,首先在网络中选取根节点和中继节点,其次叶节点将观测数据发送给上级中继节点,中继节点再将接收到来自邻居叶节点的观测数据和自身观测数据原封不动以多通道形式发送给父节点,以此类推直至根节点,最终根节点估计目标语音信号。这种方式最大的问题在于节点层数越深,发送的数据量和所需的通信带宽就越大,这对于无线智能终端设备通常是难以承载的。

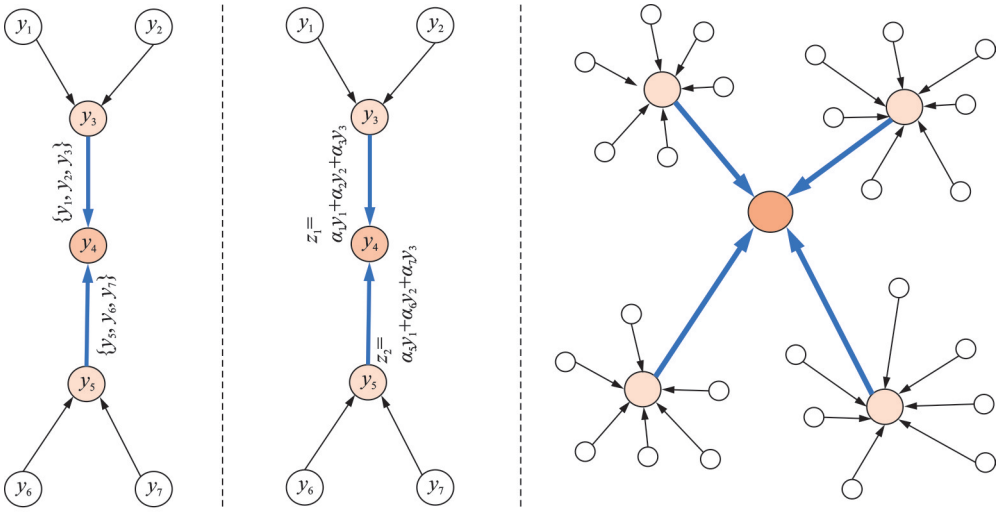

图4　分布式麦克风阵列信号融合模式

Fig.4　Signal fusion modes of distributed microphone array

局部模式:与中继模式相比,局部融合的区别在于中间节点在接收到来自邻居节点的数据后,与自身观测数据进行加权融合,融合之后的单通道信号会被发送给下个节点,以此类推直至根节点得到最终融合的单通道拾音结果,如图4(b)所示。局部融合方式能够大大降低每次传输的数据量和带宽,缺点是每次简单的加权融合都会引起信息损失。一种简单的改进策略是在局部运用延时-累加波束形成器(Delay-and-sum beamformer, DSB)[25],通过估计接收信号间的到达时间差(Time difference of arrival, TDOA)并进行时间上移位和对齐,叠加后的单通道信号质量会有一定程度上改善,DSB技术能起到初步语音增强的效果。

聚类-滤波模式：为了同时降低信息损失、数据收发能耗和通信带宽，聚类融合模式是中继融合和局部融合的折衷，如图4(c)所示。首先对所有初始节点依据空间分布进行聚类，选取每个簇的中心节点，簇内所有节点将观测数据发送给簇中心；然后簇中心节点设计局部波束形成器估计目标信号，网络对簇中心节点进行二次聚类，簇中心节点类似地将估计信号发送给新的簇中心，直至根节点完成最终滤波任务。显然这种聚类融合按树形拓扑结构收集信息，全局滤波器设计和信号估计任务分摊到各个节点簇[26-28]。这种方式估计的信号质量通常优于局部融合，但其输出结果依然是次优的，因为取决于聚类规则、最优滤波器依赖于所有节点观测信号的互统计量，而层级别簇中心节点无法获取该全局信息。

1.4 分布式信源编码

前面所述的3种信息融合方式存在资源浪费或者信息损失的弊端，这可以运用信源编码理论予以解决。以包含两个节点的无线HA为例，每个HA分别安装两个麦克风，显然麦克风信号间是高度相关并且存在大量信息冗余。信源编码可分为两个阶段：首先，利用信息论知识通过计算 $Y_{\mathrm{L},1}$ 和 $Y_{\mathrm{L},2}$ 的互信息(右侧同理)，以 $Y_{\mathrm{L},1}$ 为参考，只需要对 $Y_{\mathrm{L},2}/Y_{\mathrm{L},1}$ 进行编码，从而实现对单侧双通道信号进行辅助信息感知(Side information aware, SIA)信源压缩，如图5所示；其次考虑双耳HA间信息交互阶段，在双向通信模式下同个节点既是发送节点又是接收节点，因此任一侧HA可以获取另一侧的全部信息，根据双耳信号的互信息和香农码率-失真定理(Rate-distortion theory)[29]，在给定最大容许失真水平条件下可以计算左右耳编码器的最低码率，这相对于中继模式 $\{Y_{\mathrm{L},1}, Y_{\mathrm{L},2}\}$ 可以大大降低码率和通信带宽。辅助信息感知信源编码方法要求提供全局信号统计特性，在助听器场景下具有很高的编码效率[30-32]，对于多节点的分布式麦克风网络场景一般也适用(如典型的Distributed adaptive node-specific signal estimation (DANSE)算法)[33-36]。当无法获取全局统计特性时，可以采用辅助信息无意识(Side information unaware，SIU)信源编码技术。

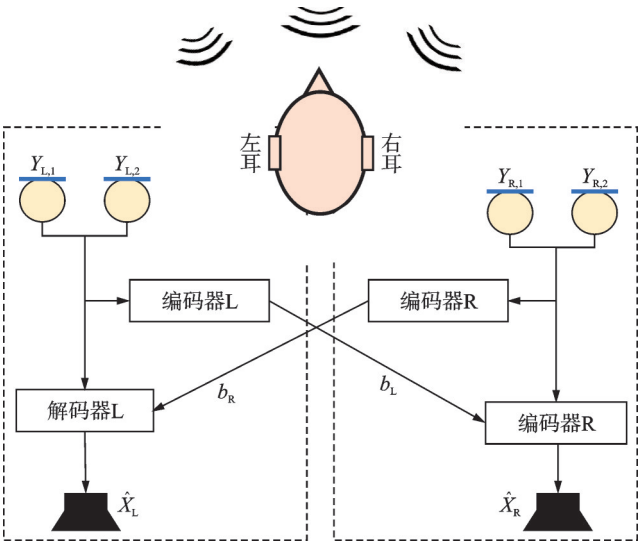

图5 双耳助听器场景下辅助信息感知信源编码

Fig.5 Side information aware source coding for binaural hearing aids SIA source coding for binaural HAs

1.5 分布式麦克风信号盲同步

分布式麦克风阵列中，各节点通常配备有独立的模/数(A/D)转换器，由于实际制造工艺、温度、湿度等差异，A/D转换器的时钟晶振频率与其标称值间存在一定漂移，实际测试时甚至可达数百ppm (Parts per million)。因此，即使所有节点标称采样频率相同，节点之间依然存在采样频率偏差(Sampling rate offset, SRO)[37]。相位是采样频率的积分，它受采样频率偏差的影响更大。SRO问题给后续语音处理算法造成了很大困扰，严重时甚至会抵消分布式麦克风阵列带来的优势。为了保证基于分布式麦克风阵列的拾音与交互系统能够稳定工作，以消除采样频率偏差为目标的时钟同步成为系统构建的前端关键步骤之一。

现有分布式麦克风阵列的时钟同步算法主要分两类，一类是通过无线通信协议在节点之间传输特定的时间戳信号，据此实现时钟同步[38-39]；另一类则是利用声信号进行时钟盲同步[40-42]。后者由于无需

任何先验,从而受到研究者们的广泛关注。目前,时钟盲同步方法通常分两步执行:首先,估计各节点之间的SRO;其次,对麦克风信号进行重采样(即补偿SRO)。受篇幅限制,下面仅讨论时钟盲同步中的SRO估计方法。

令节点i和j的标称采样率为f_s,实际采样频率分别为f_i与f_j,接收音频信号分别为$x_i[n]=x_i(t/f_i)$、$x_j[n]=x_j(t/f_j)$,其中,n为离散时间索引,$x_i(t)$为节点i未经过A/D转换的模拟信号,t表示连续时间索引。那么,信号$x_i[n]$的短时傅里叶变换(Short-time Fourier transform, STFT)可表示为

$$X_i[k,l]=\sum_{n=0}^{N-1}w[n]x_i[lN_s+n]\exp\left(-\mathrm{j}\frac{2\pi kn}{N}\right) \tag{2}$$

式中:j为虚数单位,k表示频率,l表示帧数,$w[n]$为窗函数,N为窗函数的长度,N_s为帧移。上述频域信号可近似为

$$X_i[k,l]\approx\sum_{n=0}^{N-1}w[n]x_i\left[\frac{lN_s}{f_s}+\frac{n}{f_s}-\frac{lN_s\varepsilon_i}{f_s^2}\right]\exp\left(-\mathrm{j}\frac{2\pi kn}{N}\right)=\bar{X}_i[k,l]\exp\left(-\mathrm{j}\frac{2\pi knN_s\varepsilon_i}{f_sN}\right) \tag{3}$$

式中:ε_i表示节点i的SRO,即$\varepsilon_i=f_i-f_s$,$\bar{X}_i[k,l]$表示没有采样率偏差时接收信号在时频点$[k,l]$处的STFT系数。STFT域信号模型表明,当存在SRO ε_i时,原信号在STFT域引起了线性相位漂移(Linear phase drift, LPD)。

近年来涌现出不少基于LPD模型的SRO估计方法。Markovich-Golan等[40]基于瞬时相关系数推导出SRO的闭式解。Schmalenstroeer等[41]将瞬时相关系数的相位关于频率进行加权求和,并通过多步估计SRO,显著提升了估计精度。瞬时相关系数在高频处存在相位模糊,因此上述两种方法仅通过中低频信号进行采样率偏差估计。Hu等[43]提出一种解卷绕策略以消除高频相位模糊问题,并使用全频信息进行SRO估计,进一步提升了估计精度。另外文献[43]还以分布式的方式,使各麦克风节点并行、协作地估计SRO,故无需额外的中心处理单元。尽管上述方法具有SRO的解析表达式,易于扩展为在线方法[44],但需要语音活动检测、异常值消除等前置模块的支持。此外,LPD模型表明当一对麦克风放置于不同位置时,它们接收信号的相干系数在补偿正确的SRO后得到最大值。因此,Cherkassky等[45]通过梯度上升法最大化节点间的相干系数估计SRO;Wang等[46]运用两步穷搜索策略进一步提高了最大相干系数方法的精度。另一类方法则假设经SRO补偿后的两路信号差服从零均值复高斯分布,从而应用最大似然方法估计SRO[47-48],文献[49]将该方法拓展到动态声源场景中。Masuyama等[50]采用高维复高斯分布同时对所有通道的SRO进行建模,再执行最大似然估计,当校准信号长度大于30 s时,其SRO估计误差约为10^{-4} Hz。值得注意的是,最大相干系数和最大似然估计方法均难以获得SRO的闭式解,通常需要迭代求解,因此运算复杂度较高,难以应用至在线场景中。

在线/低复杂度的SRO估计方法包括:Chinaev等[51]利用双重互相关函数对SRO进行建模,再结合二阶多项式插值进行SRO估计;该方法在文献[44]中被改进为在线模式,在文献[52]中被扩展为分布式运算模式;文献[53]提出了一种适用于动态传感器网络拓扑的分布式SRO估计方法;Chinaev等[54]进一步对双重互相关函数引入相位变换,其估计精度在实际环境中可达0.26×10^{-6};文献[55]提出一种声学相干状态度量,用于消除动态声场对SRO估计的影响;文献[42]基于LPD模型和最小二乘法,建立了联合SRO估计和声源定位方法。上述SRO盲估计方法为了保证估计精度,往往需要30 s甚至更长的声信号作为校准源。

表1中汇总了现阶段SRO盲估计代表性方法的信号处理域、问题求解、在线/离线及集中式/分布式处理模式等特点,其中在线-分布式方法有待进一步研究。

表1 分布式麦克风节点时钟同步方法概括

Table 1 Summary of clock synchronization methods for distributed microphones

方法	信号域	闭式解	在线/离线	模式
文献[40-42]	STFT	√	离线	集中式
文献[43]	STFT/时域	√	在线	集中式
文献[44]	STFT	×	离线	集中式
文献[45-50]	STFT	√	离线	分布式
文献[37,51]	时域	×	离线	集中式
文献[54-55]	STFT+时域	×	在线/离线	集中式
文献[52-53]	STFT+时域	×	在线	分布式

2 麦克风效用评估和阵列布局优化理论

在大规模分布式麦克风阵列中,不同麦克风节点由于距离目标声源的距离不同,导致其具有不同的拾音效用。由于无线设备通常具有有限的电池资源,基于分布式麦克风阵列的语音信号处理算法必须考虑节点层面和整体网络层面的能耗,其中节点能耗取决于信号传输速率和传输距离,而整体能耗还与节点数量有关。因此,分布式麦克风阵列的布局优化是平衡系统成本和拾音质量的关键,成为近年来国内外的研究热点之一,其中衡量麦克风节点效用是核心。

2.1 基于信号时频特征的麦克风效用估计

研究表明,信号波形特征(如过零率、统计矩、波形熵)和频谱特征(如幅度谱偏度、峰度、斜率、功率平坦度、幅度平坦度)可用于重构多麦克风信号的幅度平方相干(Magnitude square coherence, MSC)系数。Gunther等[56-58]提出了基于语音信号时频特征的麦克风效用估计方法,如图6所示。

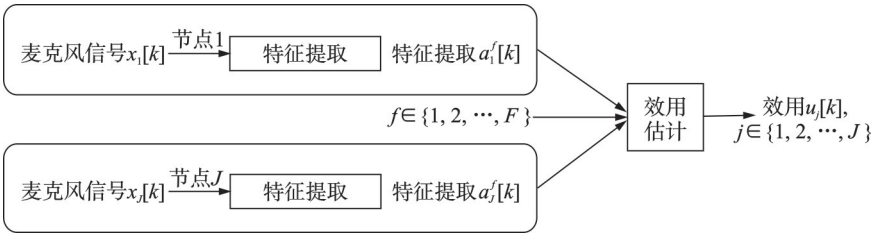

图6 基于时频特征的麦克风效用估计模型

Fig.6 Microphone utility estimation model using time-frequency features

该方法首先对观测音频信号进行分块操作,令第j个麦克风第k块信号矢量表示为$\boldsymbol{x}_j[k]=[x_j[kM],x_j[kM+1],\cdots,x_j[kM+L-1]]^{\mathrm{T}}$,其中$M$和$L$分别表示块移和块长,可以采用频率平均技术($n$为频率索引,$L$个频点)估计窄带MSC系数为

$$\gamma_j=\frac{1}{L}\sum_{n=1}^{L}\left|\frac{\Phi_{s,x_j}[k,n]}{(\Phi_{s,s}[k,n]\Phi_{x_j,x_j}[k,n])^{1/2}}\right|^2 \tag{4}$$

式中:Φ_{s,x_j}和$\Phi_{s,s}$分别代表互功率谱和自功率谱密度函数。然后,针对每一信号块提取声学特征,令第k块的第f个特征矢量表示为$\boldsymbol{a}^f[k]\in\mathbb{R}^J$,向量$\boldsymbol{a}^f[k]$与其自身的外积表征了特征$f$的通道间瞬态相关性,在所有观测块上取时间平均就可以估计出皮尔逊相关系数(Pearson correlation coefficient, PCC)矩阵$\boldsymbol{C}_j[k]$,该矩阵共包含F列,每列刻画了相应特征表征的通道间相干性,每行刻画了相应麦克风相对于

参考通道j的相干性,意味着矩阵$\boldsymbol{C}_j[k]$的主特征向量暗含了其他通道相对于参考麦克风j的相似性。令$\boldsymbol{r}_j[k]$和$\boldsymbol{v}_j[k]$分别表示主左奇异向量和主右奇异向量,分别刻画了每个通道和每个声学特征对矩阵$\boldsymbol{C}_j[k]$结构的贡献,类似地,分别选择其他麦克风作为参考,可估计出PCC矩阵集合$\{\boldsymbol{C}_1[k],\boldsymbol{C}_2[k],\cdots,\boldsymbol{C}_J[k]\}$和相应的相似性矢量集合$\{\boldsymbol{r}_1[k],\boldsymbol{r}_2[k],\cdots,\boldsymbol{r}_J[k]\}$,进而按式(5)构造出所有麦克风间相似性矩阵。

$$\boldsymbol{R}[k]=\left[\frac{\boldsymbol{r}_1[k]}{[\boldsymbol{r}_1[k]]_1},\frac{\boldsymbol{r}_2[k]}{[\boldsymbol{r}_2[k]]_1},\cdots,\frac{\boldsymbol{r}_J[k]}{[\boldsymbol{r}_J[k]]_1}\right]\in\mathbb{R}^{J\times J} \tag{5}$$

采用无向图$\mathcal{G}(\mathcal{V},\mathcal{E})$对麦克风网络进行建模,顶点集$\mathcal{V}$包含所有麦克风节点,边集合$\mathcal{E}$包含所有麦克风对之间的相似性。邻接矩阵的计算方式如下

$$\boldsymbol{W}(k)=0.5\times\left(\boldsymbol{R}[k]+\boldsymbol{R}^{\mathrm{T}}[k]\right) \tag{6}$$

每个顶点与其相邻顶点的连接权重进行求和可以得到顶点的度(Degree),所有顶点的度构成的对角矩阵即为度矩阵$\boldsymbol{D}[k]$。令J-维离散矢量$\boldsymbol{u}[k]$表示麦克风的显著性,无向图$\mathcal{G}$的顶点显著性分析等价为谱图分割问题

$$\{\mathcal{S},\bar{\mathcal{S}}\}=\min_{\boldsymbol{u}[k]}\frac{\boldsymbol{u}^{\mathrm{T}}[k]\left(\boldsymbol{D}[k]-\boldsymbol{W}[k]\right)\boldsymbol{u}[k]}{\boldsymbol{u}^{\mathrm{T}}[k]\boldsymbol{D}[k]\boldsymbol{u}[k]}\quad\text{s.t. }\boldsymbol{u}^{\mathrm{T}}[k]\boldsymbol{D}[k]\boldsymbol{1}_J=0 \tag{7}$$

可以证明$\boldsymbol{u}[k]$为拉普拉斯矩阵$\boldsymbol{L}[k]=\boldsymbol{I}_J-\boldsymbol{D}^{-1}[k]\boldsymbol{W}[k]$的特征向量,矩阵$\boldsymbol{L}[k]$的次最小特征值所对应的特征向量(菲德勒矢量)即为麦克风效用矢量$\boldsymbol{u}[k]$。在给定最大容许麦克风数量条件下,可以直接选择$\boldsymbol{u}[k]$较大元素对应的麦克风构成具有较大拾音效用的分布式麦克风子阵列,其他节点设备可关闭以节省能耗。

2.2　基于节点增删的麦克风效用估计

针对语音增强任务,Naraganan等[59-61]提出了基于节点删减的麦克风效用估计方法,不失一般性以语音失真加权多通道维纳滤波器(Speech distortion weighted multichannel Wiener filter, SDW-MWF)[62]设计为例,其目标函数如下

$$J(\boldsymbol{w})=\mathbb{E}\left\{\left|x_1-\boldsymbol{w}^{\mathrm{H}}\boldsymbol{x}\right|^2\right\}+\mu\mathbb{E}\left\{\left|\boldsymbol{w}^{\mathrm{H}}\boldsymbol{v}\right|^2\right\} \tag{8}$$

式中:x_1代表第1个麦克风采集的干净信号成分,$\boldsymbol{x}$为某时频点上所有干净信号成分构成的列向量,$\boldsymbol{v}$为所有麦克风采集的噪声分量构成的列向量,H表示共轭转置。运用拉格朗日乘子法,滤波器为

$$\boldsymbol{w}=\frac{\boldsymbol{R}_{vv}^{-1}\boldsymbol{R}_{xx}\boldsymbol{e}_1}{\mu+\mathrm{tr}\left(\boldsymbol{R}_{vv}^{-1}\boldsymbol{R}_{xx}\right)} \tag{9}$$

式中:$\boldsymbol{R}_{xx}$和$\boldsymbol{R}_{vv}$分别代表干净信号和噪声分量的互相关矩阵,$\boldsymbol{e}_1$为标识向量(第1个元素为1,其余均为0)。可以证明,SDW-MWF滤波器的输出信噪比随着μ增大而增大,随着矩阵$\boldsymbol{R}_{xx}$的秩增大而减小,但是选择合适的权衡因子μ和秩对改善增强信号的可懂度有利[63]。

进而考虑删除分布式麦克风阵列中第m个节点,按照上面全节点方式设计新的滤波器$\boldsymbol{w}_{-m}$,其代价函数与式(8)的差值定义为节点m的效用,即

$$U_m=J_{-k}(\boldsymbol{w}_{-m})-J(\boldsymbol{w}) \tag{10}$$

很明显,移除节点m引起的代价函数损失越小意味着其效用越小,因为该节点对降噪的贡献越小。除了考察代价函数,也可以计算滤波之后的信噪比变化,移除某节点引起的信噪比下降越小意味着该节点的效用也越小。

移除节点可以理解为逐渐缩小阵列的考察范围,在增大分布式麦克风阵列规模时可以使用正向逐渐增加节点的方式计算新节点的效用,定义如下

$$U'_m = J(\boldsymbol{w}) - J_{+k}(\boldsymbol{w}_{+m}) \tag{11}$$

基于节点增删操作获得的麦克风效用估计适用于语音增强任务,对其他任务的有效性有待验证。利用麦克风效用可以从大规模分布式节点中选择效用高的麦克风子集用于语音增强,但该方法的时间复杂度很高,需要重复设计滤波器和计算每个节点效用,直至选择了规定数量的麦克风或满足了期望语音增强性能[64]。

2.3 基于阵列能耗优化的麦克风效用估计

鉴于分布式麦克风阵列能耗与麦克风数量和传输速率密切相关,近年来涌现出不少面向大规模无线声传感器网络的麦克风子集选择和传输速率优化方法。此类方法显式地优化分布式麦克风阵列布局,隐式地评估麦克风节点效用。

(1)麦克风子集选择。针对单一目标语音增强问题,文献[65]提出了基于麦克风子集选择的最小方差无失真响应(Minimum variance distortionless response, MVDR)波束形成方法,其核心思想是最小化阵列节点总体能耗、约束输出噪声功率,优化问题描述如下

$$\min_{\boldsymbol{w}_p,\boldsymbol{p}\in\{0,1\}^M} \left\| \mathrm{diag}(\boldsymbol{p})c \right\|_1 \text{ s.t. } \quad \boldsymbol{w}_p^{\mathrm{H}}\boldsymbol{R}_{vv,p}\boldsymbol{w}_p \leqslant \frac{\beta}{\alpha}, \quad \boldsymbol{w}_p^{\mathrm{H}}\boldsymbol{a}_p = 1 \tag{12}$$

式中:$\boldsymbol{p}$代表标识麦克风节点是否被选择的布尔矢量,$\boldsymbol{a}$为声源到麦克风阵列的声学传递函数(Acoustic transfer function, ATF),c代表麦克风到中心节点的传输能耗矢量,β代表使用所有麦克风情况下得到的最小输出噪声功率,$0<\alpha\leqslant 1$用于控制期望输出噪声功率,线性约束条件$\boldsymbol{w}_p^{\mathrm{H}}\boldsymbol{a}_p=1$用于控制滤波器在目标声源方向的响应无失真。麦克风子集优化问题式(12)属于组合优化问题,通过分析拉格朗日函数不难发现经典的MVDR波束形成器是其最优滤波器,那么式(12)可以简化为纯麦克风选择问题。利用凸松弛(Convex relaxation)技术、将$\boldsymbol{p}$放松至0~1区间连续变量,该问题可以转化为半正定规划(Semi-definite programming, SDP)问题,然后使用成熟的凸优化工具求解(如SeDuMi、CVX)。最后,需要对连续结果$\boldsymbol{p}$进行规整,从而得到二值分布的麦克风选择状态。实验结果表明,该方法在取得同等语音增强性能的条件下,相较于节点增删方法[64]的网络能耗更低。另外,连续变量$\boldsymbol{p}$也可以解释为麦克风节点对降噪任务的贡献,即效用。更进一步,文献[66]从理论上分析了ATF参数估计误差对麦克风子集优化的影响;由于式(12)依赖于每个频率点信息,其选择结果需要在频率间频繁切换,文献[67]通过约束宽带信噪比建立了宽带频率不变的麦克风子集优化方法。对于基于麦克风子集选择的MVDR波束形成方法所需的归一化ATF参数(即相对ATF)可以采用协方差差分或白化(Covariance subtraction/whitening)方法估计[68-70],参考麦克风可使用文献[71]的方法优化。

(2)传感器速率优化。麦克风子集选择方法虽然能够大大降低传感器数量,但每个选中的传感器会选择规定的速率进行数据量化编码和传输。式(1)表明了数据传输能耗与速率呈指数关系,因此优化传感器速率分配也可以节约网络能耗。文献[72]通过最小化总体网络能耗、约束输出噪声功率的方式,建立了传感器速率优化方法如下

$$\begin{aligned} &\min_{\boldsymbol{w},b} \sum_{j=1}^{J} d_j^2 N_j\left(2^{2b_j}-1\right) \\ &\text{s.t. } \boldsymbol{w}^{\mathrm{H}}\boldsymbol{R}_{v+q}\boldsymbol{w} \leqslant \frac{\beta}{\alpha},\ \boldsymbol{\Lambda}^{\mathrm{H}}w = \boldsymbol{f} \\ &b_j\in\mathbb{Z}_+, \quad b_j \leqslant b_0,\ \forall j \end{aligned} \tag{13}$$

式中:b_j、d_j和N_j分别代表节点j分配的传输速率、节点到中心计算机的传输距离和传输信道噪声功率,$\boldsymbol{R}_{v+q}$为麦克风阵列观测噪声叠加均匀量化噪声的噪声互相关矩阵,b_0为最大允许分配的速率(单位:比

特/样本),线性约束条件用以限制滤波器$\boldsymbol{w}$在目标方向上的响应。分析拉格朗日函数可以发现经典的线性约束最小方差(Linearly-constrained minimum variance, LCMV)波束形成器[73]是式(13)的最优滤波器(单一线性约束条件下MVDR滤波器是其最优滤波器)。代入LCMV滤波器表达式可以降低优化变量数目,但该问题依然是整数非凸优化问题。运用凸松弛技术、并将$\boldsymbol{b}$松弛为连续变量,式(13)也可以转化为典型的SDP问题,最后整数速率需要利用规整技术获得。本质上,速率也可以理解为麦克风效用,因为效用高的传感器分配的速率通常较高,速率分配问题是麦克风子集选择问题的更一般形式[74],0-速率传感器即未被选中。另外,文献[75]中针对分布式LCMV波束形成任务,运用主对偶乘子法(Primal-dual method of multipliers, PDMM)[76]解决了分布式传输速率优化问题。

值得注意的是,若(13)的目标函数包含麦克风选择变量,那么式(13)等价于联合麦克风子集选择和速率优化问题[77]。为了更直观地分析麦克风子集选择和速率优化算法暗含的麦克风效用评估功能,这里设置8 m × 6 m × 3 m的三维房间,包含1个目标说话人和2个干扰声源,混响时间$T_{60}=200$ ms,随机放置$J=100$个麦克风节点,麦克风子集选择和速率优化算法仿真结果如图7所示,其中蓝色点代表选中的麦克风。可以很明显看出靠近声源和中心节点的麦克风被选中的概率更高、被分配的速率更高,代表其麦克风效用更高。靠近目标说话人附近的麦克风通常信噪比较高,对语音增强有益;靠近噪声源的麦克风虽然信噪比较低,但很好地记录了噪声信息,对噪声消除有用;中心节点周围的麦克风传输距离较短,在传输能耗方面更有优势。

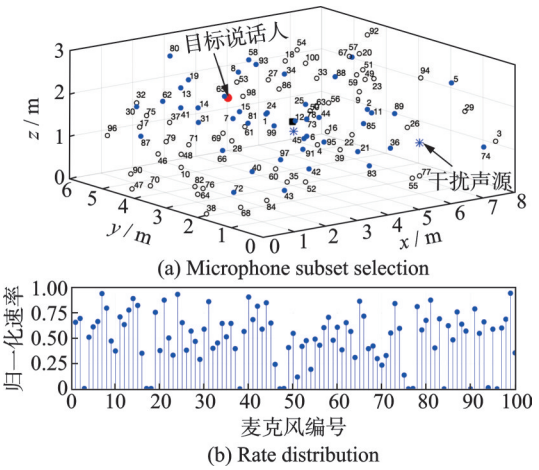

图7　三维房间分布式麦克风阵列传感器子集选择和速率优化仿真结果

Fig.7　Simulation results of microphone subset selection and rate optimization for distributed microphone arrays in a 3D room

3　分布式麦克风阵列应用方法

本节将结合语音增强、声源定位、说话人识别、语音识别、说话人日志和音频场景分析等下游任务阐述分布式麦克风阵列在真实声场景中的应用方法。

3.1　基于分布式麦克风阵列的语音增强

语音增强旨在抑制噪声成分、保留目标语音信号,是前端提升拾音质量的关键步骤,已成为噪声场景下语音交互系统不可或缺的模块之一。基于麦克风阵列波束形成的语音增强技术通常表现优于单麦克风方法,依赖于波达方向(Direction of arrival, DOA)、协方差矩阵、参考麦克风、ATF等参数。由于分布式麦克风阵列的布阵和通道数量存在随机性,动态随机的拓扑结构也使得分布式麦克风阵列在现实中很难获得精确的空间指纹,使得传统麦克风阵列波束形成技术并不能直接应用于分布式语音增强任务。因此,近年来涌现出许多基于分布式麦克风阵列的语音增强方案,大体可分为基于子阵列选择的分布式语音增强、基于波束形成的分布式语音增强、基于深度学习的分布式语音增强和基于特征值分解的分布式语音增强4类,下面将具体说明。

(1)基于子阵列优化的语音增强。在大规模分布式麦克风阵列系统中所有传感器的能耗代价是非常昂贵的,选择对语音增强任务贡献度高的麦克风子集合、排除效用低的麦克风能够大大节约数据传输能耗和计算复杂度。因此,在限制麦克风数量或拾音信号质量的条件下优化子阵列具有重要的理论意义和经济价值,该问题理论上均可写为下面的约束优化问题

$$\min_{\mathcal{S}} f(\boldsymbol{w},\mathcal{S}) \quad \text{s.t. } g(\boldsymbol{w},\mathcal{S})\leqslant\alpha,\ |\mathcal{S}|\leqslant K \tag{14}$$

式中：$\boldsymbol{w}$表示语音增强滤波器，$\mathcal{S}$为选中的麦克风集合，$f(\boldsymbol{w},\mathcal{S})$和$g(\boldsymbol{w},\mathcal{S})$分别代表能耗和增强性能度量函数，$\alpha$为性能阈值，$K$为最多允许的麦克风数量。可以证明，式(12)和式(13)均为式(14)的特例，注意该问题本质上属于非凸组合优化问题。

文献[65]通过最小化总功耗和约束输出噪声功率，建立了基于麦克风子集选择的MVDR波束形成方法，文献[66]证实了限制被选传感器的数量和最大化输出SNR实现子阵列优化的可行性，这两种方法可直接拓展至DSB、LCMV、SDW-MWF等其他滤波器。值得注意的是麦克风子集选择方法通常包含布尔约束条件，需要采用凸松弛或贪婪搜索策略来寻找次优解，其计算复杂度随麦克风数量呈立方次上升。文献[78]基于无线通信理论中能量效率，针对DSB语音增强任务，通过最大化输出SNR与能耗比值，运用丁克尔巴赫求解器建立了一种快速的子网络选择方法。文献[77]提出了一种联合速率分配和子阵列优化方法，其中目标函数式(14)为$\mathcal{S}$和比特率变量的复合函数，该问题属于混合半定和双线性规划问题，需要采用两步优化策略进行求解。文献[79]将整数非线性规划问题转化为多个子优化问题，每个子优化问题都采用最速下降(Steepest decent, SD)法求解，进而提出了一种贪婪搜索策略来快速选择传感器，并给出了更适用于分布式传感器网络的分布式SD算法。文献[80]通过只激活一个信息子网络，并引入子网络连通性约束条件，提出了一种近似最优的贪婪方法，在每次迭代中删除一个传感器直到子网络规模达到预设值。文献[81]基于伪相干系数选择子阵列和参考麦克风，推导了MVDR、SDW-MWF、LCMV等多种波束形成器，建立了一个适用于自组织麦克风阵列的语音增强框架。

(2)基于分布式信号处理的分布式语音增强。文献[82]运用PDMM优化算法最大化信号干扰噪声比(Signal-to-noise interference ratio, SINR)，提出了不局限于特殊图拓扑结构的分布式信号子空间滤波方法。本质上，PDMM[76]是交换乘子方向法(Alternating direction method of multipliers, ADMM)的变体，其优势在于利用双向数据交换节约通信次数。ADMM、PDMM和随机流言(Randomized gossip)[83]是最常用的3种针对约束优化问题的分布式算法，具体算法原理不在此讨论，感兴趣读者请参阅文献[76,83-85]及其他相关文献。这些分布式信号处理算法均可用于语音增强任务，例如文献[25,86]建立了基于随机流言策略的分布式DSB方法，对大型分布式麦克风阵列和动态环境具有良好的鲁棒性和可扩展性。

针对理想全连通无线传感器网络，Bertrand等[33]在提出了著名的DANSE算法，其设计思想是以分布式方式在每个节点上估计一个节点特定的信号。文献[34]将其扩展到任意节点数量和多目标说话人场景，并支持异步节点更新模式；文献[87]将DANSE算法应用于自适应分布式降噪任务，在噪声平稳条件下其结果收敛于集中式MWF；文献[35]将DANSE算法扩展到树形拓扑的WASNs，表明了其通信带宽和计算能力的可扩展性；随后他们又建立了分布式LCMV波束形成语音增强方法[88]，Markovich-Golan等[89]针对全连接WASN提出了广义旁瓣抵消器(Generalized sidelobe canceller, GSC)的递归分布式版本；文献[90]将DANSE算法推广至混合拓扑结构的WASN。文献[91]基于可访问的自顶向下框架分析了3种基于LCMV的最优分布式算法的关系，并解释了如何扩展到可剪枝至树形拓扑的WASN。针对WASN中集中求解所有节点特定信号的复杂度过高问题，文献[92]提出了一种数据驱动的信号流方法，支持动态拓扑结构。这些方法本质上都是基于DANSE实现的。

文献[93]提出了基于广义线性坐标下降的分布式MVDR波束形成器，是分布式DSB[25,86]更一般的形式。文献[94]建立了基于扩散自适应范式的分布式MVDR滤波器，支持时间上连续自适应滤波。文献[95]提出了一种LCMV的分布式重构算法，其中最优滤波器的输出可以在每个节点计算，从而不需要在网络内共享原始数据。文献[96]提出了基于双向ADMM的稀疏MVDR波束形成技术，双向

ADMM正是PDMM的雏形。文献[97]利用全局互相关信息对参数化MWF进行改进,使用标签矩阵与互相关矩阵的点积建立了分布式MWF语音增强方法,支持通信开销与增强性能之间的灵活折衷(Trade-off)。

(3)基于深度学习的分布式语音增强。近年来,基于深度学习的麦克风阵列语音增强迅速发展,深度神经网络(Deep neural network, DNN)模型也促使估计相对声学传递函数、语音协方差矩阵等参数取得了长足进步,对分布式麦克风阵列语音增强任务也表现出很好的适用性。

文献[98]针对空间无约束分布式麦克风阵列,提出了基于DNN时频掩码估计的分布式语音增强方法,使用局部维纳滤波后的压缩信号作为空间信息用于目标声源和噪声估计。当空间信息通过STFT系数隐式地提供给DNN时,现实中麦克风数量的变化会导致模型输入信号维度发生变化,从而影响语音增强的效果。针对该问题,可以在输入通道之间使用共享参数;考虑到输入通道的差异性(分布式麦克风具有不同的拾音效用),因此需要使用注意力机制赋予通道权重,从而有效融合多通道信息。文献[99]参考传统神经波束形成器训练策略设计了深度ad-hoc波束形成方法,其核心模块包括单通道语音增强、跨通道特征拼接、SNR和通道权重估计,然后依据通道权重确定目标说话人信息量高的稀疏麦克风子集,最后进行时间对齐以输出单通道增强语音,其核心思想与注意力机制类似。

考虑到分布式麦克风阵列布局的可变性,确保训练的DNN对不同麦克风阵列排布的泛化能力依然是现阶段多麦克风语音增强领域的热点和难点课题之一。麦克风阵列语音增强网络的泛化需要解决两个问题:①优化选择参考通道,②提取空间特征,即包含声源位置、距离等空间信息,这些特征应通过多麦克风信号的相关性来获取,而不是麦克风阵列的物理排列。文献[100]基于通道填充、通道卷积、通道配对等策略提出了一种窄带深度滤波器,其基线系统是长短时记忆网络(Long short-term memory, LSTM),训练数据包括116种虚拟仿真阵列(包含麦克风数量2~8个不等),显著提高了DNN对阵型的泛化能力,但该方法对真实噪声场景、测试麦克风数量多于训练数量等条件的有效性缺乏验证。

(4)基于子空间分解(Subspace decomposition)的分布式语音增强。鉴于波束形成方法严重依赖于阵列拓扑,因此研究拓扑参数独立的分布式麦克风阵列语音增强方法更有实用价值。Neo等[101]提出了多项式特征值分解(Polynomial eigenvalue decomposition, PEVD)方法,它充分考虑了麦克风信号的时频相关性,利用空时相关矩阵估计期望信号子空间。令空时相关矩阵为

$$\boldsymbol{R}_{xx}(\tau)=\mathbb{E}\left[\boldsymbol{x}(n)\boldsymbol{x}^{\mathrm{H}}(n-\tau)\right] \tag{15}$$

式中:$\boldsymbol{x}(n)$为Q-子阵列多麦克风信号矢量,τ为时间延迟。空时变换矩阵经过z变换后,它的PEVD写作

$$\boldsymbol{R}_{xx}(z)=\boldsymbol{U}(z)\boldsymbol{\Lambda}(z)\boldsymbol{U}^{\mathrm{P}}(z) \tag{16}$$

式中:$\boldsymbol{U}(z)$为特征向量多项式矩阵,$\boldsymbol{\Lambda}(z)$包含特征值,$[\cdot]^{\mathrm{P}}$为仿Hermitian算子。

最终增强后的语音信号通过主特征向量波束形成得到

$$Y^{\mathrm{PEVD}}(z)=\boldsymbol{u}_1^{\mathrm{P}}(z)\boldsymbol{x}(z) \tag{17}$$

设置矩阵$\boldsymbol{R}_{xx}$的秩,即低秩分解,可构造rank-r滤波器,其中MVDR滤波器是rank-1特例,MWF是满秩特例[63,71,102]。

由于语音信号在时间和频率上呈现周期性,PEVD算法能够捕获语音信号的周期特性,通常具有更好的增强潜力。因此,文献[103]将PEVD算法拓展至分布式麦克风阵列语音增强任务。为了降低PEVD方法的计算量,每个Q-子阵列先进行节点层面局部波束形成产生Q个融合信号,然后运用PEVD算法生成单通道输出信号。该方法避免了跨阵列波束形成操作,不需要进行无线设备同步和RTF估计。针对动态随机的分布式麦克风阵列拓扑结构导致的空间指纹难以估计这一挑战性问题,文献[104]使

用广义特征值分解(GEVD)技术在全连接网络中估计信号协方差矩阵,进而在LCMV波束形成中使用投影算子估计信号/噪声子空间[105],其核心想法是在麦克风数量大于所需子空间维数的节点中局部估计协方差矩阵,从而减少节点间通信频次。

3.2 基于分布式麦克风阵列的声源定位

声源定位作为音频信号处理领域中备受关注的研究任务之一,广泛应用于军事、公共安全监控、人机交互等领域,常常也是语音增强、声源分离等语音处理任务的前端步骤。相较于传统麦克风阵列,分布式麦克风阵列具有更大的空间覆盖度,具备更高的声源定位精度。依据定位特征线索,分布式麦克风阵列的声源定位方法大致分为如下5类。

(1)基于到达时间(Time of arrival, TOA)的声源定位。在已知声源波形条件下,可以根据它与麦克风接收信号的互相关函数能估计声源到麦克风节点的TOA。通常TOA等于飞行时间(Time of flight, TOF)、声源发射时刻和节点起始采样时刻的三者之和,其中TOF隐含着源到节点的距离信息。假设声源发射时刻及节点起始采样时刻已知,则TOA只依赖于TOF,进而声源位置可通过三角定位法得到[106-108],但该方法涉及矩阵求逆运算,其计算复杂度较高。Zou等[109]利用迭代策略,避免了矩阵求逆运算,然而声源发射时刻及节点起始采样时刻在实际中却难以获取,往往需要与声源位置一同估计。Cobos等[110]仅假设声源发射时刻已知,在节点内运用累积求和策略估计声源发射时刻,然后求解半双曲面函数进行声源定位;他们还应用BeepBeep策略同步估计声源发射时刻和节点起始采样时刻,以及峰值匹配算法提升定位精度[111]。由于TOA估计要求声源波形已知,基于TOA的声源定位方法仅适用于主动定位场景。笔者结合随机流言算法[84],运用结构化总体最小二乘法实现了节点起始采样时刻、TOA和节点位置的同步估计[112]。

(2)基于到达时间差(TDOA)的声源定位。在被动定位场景中,即声源波形未知,可基于麦克风节点接收的语音信号计算每对节点关于声源的TDOA,避免了对声源发射时刻的依赖。当所有麦克风节点共享同一个采样设备时,Chan等[113]引入额外的距离变量,将原有的非线性TDOA方程转化为线性方程组,得到了声源位置的闭式解,但该方法的定位精度随着背景噪声增大而显著下降,噪声鲁棒性可以使用最大最小化技术进行一定程度上改善[114]。当各节点配备独立时钟设备时,TDOA估计会受到起始采样时刻的干扰,因此涌现出不少基于节点内麦克风间TDOA估计的声源定位方法[115-118]。此外,室内混响是影响TDOA估计准确率的重要因素,TDOA异常值修复方法包括基于代价函数的后验评估[119-121]、广义互相关函数峰值选择[119]和几何约束[121-122]等。近年来也涌现出一些基于深度学习的TDOA声源定位方法,其定位精度较高、但计算复杂度也较高[123-125]。上述方法要求所有麦克风节点将音频数据传输至中心处理单元运算,这造成了严重的通信能耗和带宽负担。基于TDOA的分布式声源定位方法[126-127]仅通过相邻节点间的局部数据通信,各节点并行协作地完成声源定位,避免了中心处理单元和较大的通信负担。因此,分布式声源定位成为当前热门的研究方向之一。

(3)基于能量(energy)的声源定位。除TOA或TDOA等时间线索,麦克风接收信号的能量也可用于声源定位。能量信息不受声源发射时刻、节点起始采样偏差等时间参数的干扰。在给定路径损失系数条件下,Meesookho等[128]提出了一种基于能量的加权最小二乘方法,并给出了克拉美罗下界(Cramér-Rao lower bound, CRLB)。然而,实际场景中路径损失系数往往与温度、湿度等环境参数相关,难以获取其精确数值,更好的做法是将声源位置与路径损失一同估计。例如,Wang等[129]提出一种交替估计策略,同时估计声源位置和路径损失系数。当路径损失系数未知时,总是无法避免非凸目标函数,导致问题求解难度增加[130]。因此,Vaghefi等[131]则通过使用SDP技术来对非凸代价函数进行松弛,类似地,Hu等[132]提出了基于SDP的鲁棒坐标估计器。真实环境噪声、干扰源和混响等因素会影响能量估计的精度,从而限制了此类方法的声源定位精度。

(4)基于到达方向(DOA)的声源定位方法。若节点配备紧凑型麦克风阵列,则可根据节点内麦克风信号估计声源的DOA[133-135]。基于DOA的最大似然估计往往能得到声源位置的闭式解,具有低复杂度、易于实现等优势[136-137]。由于无法得到距离相关信息,此类方法的定位精度有限。对此,Doğançay[138]提出了基于总体加权最小二乘算法的声源定位方法,Wang等[139]提出了基于位置惩罚的最大似然估计方法。尽管这两种方法显著提升了声源定位精度,但均假设所有麦克风节点的位姿已知。现实中麦克风位置需要运用分布式节点自定位方法进行几何校准,该过程不可避免会存在节点位姿误差。Wang等[140]对目标函数引入额外约束条件,提升了节点位姿误差的鲁棒性。相较于TOA、TDOA、能量等特征,DOA估计无需节点间操作,仅通过节点内少量音频数据就能获得较高的估计精度。因此,基于DOA的方法往往具有良好的实时性。

(5)基于相位变换加权的可控响应功率(SRP-PHAT)的声源定位方法。此类方法也被称为基于波束形成的声源定位,旨在将波束形成器具有最大输出功率的位置作为声源位置[141-143]。通常这类方法需将空间进行网格划分,再通过网格搜索确定声源位置[144],定位精度与网格数量密切相关,该算法的高时间复杂度问题在文献[145-146]中得到有效缓解。Salvati等[147]利用最大池化操作来减小噪声和混响的影响,注意到SRP-PHAT亦可作为DNN的输入特征来进行声源定位[148-149]。为减少通信宽带,Çakmak等[150]提出了一种基于SRP-PHAT的分布式声源定位方法,展现出良好的噪声和混响鲁棒性,但运算复杂度依然较高。文献[151]针对噪声和混响条件建立了基于观测音频特征向量到声源位置的贝叶斯映射和麦克风对协方差矩阵流形高斯随机过程的半监督定位方法,支持流式音频输入,具有较好的实时性。

表2中总结了部分已有分布式麦克风阵列声源定位方法所使用的声学特征、问题求解方式、时间参数(声源发射时刻或节点起始采样偏差)估计和集中式/分布式处理模式等特点。受篇幅限制,本节仅涉及固定声源定位问题,将上述方法与声源运动方程相结合,以概率滤波的方式可以处理实时声源跟踪问题[152-153]。另外,结合视频等其他模态信息的多模态声源定位往往能获得更好的定位性能。

表2　基于分布式麦克风阵列的声源定位方法汇总

Table 2　Summary of sound source localization methods based on distributed microphone arrays

方法	声学特征	求解方法	时间参数估计	处理模式
文献[107]	TOA	闭式求解法	×	集中式
文献[108-109]		优化方法	√	
文献[110-111]				
文献[112]		随机流言法	√	分布式
文献[113]	TDOA	闭式求解法	×	集中式
文献[114]		优化方法	√	
文献[115-122]				
文献[123-125]		机器学习	×	分布式
文献[126,127]		优化方法		
文献[128]	Energy	闭式求解法	√	集中式
文献[129-132]		优化方法		
文献[136-138,140]	DOA	闭式求解法	√	集中式
文献[139]		优化方法		
文献[141-147]	SRP-PHAT	网格搜索	√	集中式
文献[150]				分布式
文献[148-149,151]		机器学习		集中式

3.3 基于分布式麦克风阵列的说话人识别

说话人识别技术通过分析语音信号来确定说话人的身份，主要分为说话人确认(Speaker verification)和说话人辨识(Speaker identification)。前者是验证语音是否来自某特定说话人，后者是从一组候选说话人中确定语音的所属者。说话人识别技术在智慧家居、监控系统、会议等场景有着广泛的应用。然而，由于环境噪声、混响和多路径效应的影响，远场说话人识别面临诸多挑战。分布式麦克风阵列由于其多设备协同拾音和处理声源信号的能力，提供更好的空间解析能力和抗噪性能，有望提升噪声和远场条件下说话人识别系统的准确性。

基于麦克风阵列的说话人识别技术依赖于模块化的信号预处理、特征提取、特征融合、分类识别等步骤，如图8所示。分布式麦克风阵列空间随机分布的设备能捕捉到不同程度背景噪声和回声，这有益于音频数据预处理。需要指出的是，波束形成是数据预处理的常用技术之一，通过加权和合并来自不同麦克风的信号，以抑制干扰源和提升目标说话人信号的信噪比；通过估计麦克风间回声路径的时延，可以调整信号时序以消除回声干扰。例如，McCowan等[154]使用自适应波束形成器和超指向波束形成作为前端处理模块，有效提升了说话人识别准确率，Xu等[155]通过改进DSB算法对近场数据进行补偿，能够进行稳定的实时说话人识别且提供良好的说话人跟踪性能。

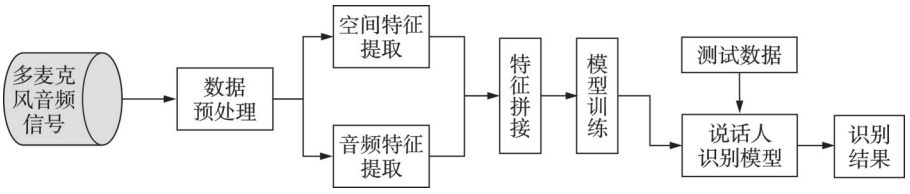

图8 基于麦克风阵列的说话人识别流程

Fig.8 Pipeline of speaker recognition based on microphone arrays

分布式麦克风阵列采集的多通道语音信号同时包含了目标说话人的内容信息和空间位置信息，其中梅尔频率倒谱系数(MFCC)是最常用的表征说话人的语音特征之一[156]。常用的说话人空间特征包括传统的通道间相位差、TDOA、电平差、广义互相关函数，以及最新的正余弦方向函数、空间一致性等[157-158]，通常特征组合越丰富，目标说话人表征准确度越高。

GMM-HMM是经典说话人识别框架，其核心包括3步：使用训练数据拟合每个说话人的GMM；结合GMM与HMM建模语音信号的时序特性；利用最大似然估计推理测试语音信号与每个模型的匹配度以确定说话人身份[159-161]。Anguera等[162]利用HMM将麦克风阵列的到达角信息与波束形成的语音频谱联合起来，显著提高了识别准确率，尤其适用于参与者静止或小幅运动的情况。孙磊[163]基于空间聚类的时频掩码重估计在无任何先验知识的条件下，利用神经网络建立了更新的说话人识别模型。

传统说话人识别方法依赖于手工设计的特征，近年来兴起的端到端方法则能够自动从数据中学习到更为稳健的声学特征。端到端方法将整个说话人识别任务作为一个整体优化，从输入的语音信号到最终的识别结果之间不再需要多个独立的处理步骤，从而避免了复杂的模型设计和潜在的错误传播。例如，文献[164]在大约100 m^2的空间中分布式放置了40个麦克风，通过采集重放的Librispeech数据集音频，生成了大规模自组织麦克风音频数据集Libri-adhoc40，并利用自注意力和图注意力机制对多通道麦克风信号进行深度融合，在说话人确认任务上有效降低了等错误率[165-166]。

分布式麦克风阵列近年来在说话人识别任务上取得了显著进展，但在复杂环境下的鲁棒性、计算资源限制等方面仍面临挑战，未来结合视频、文本等多模态信息有望进一步提升说话人识别性能和丰富其应用场景[167]。

3.4　基于分布式麦克风阵列的语音识别和说话人日志

相较传统规则型麦克风阵列或单麦克风系统,分布式麦克风阵列在自动语音识别(Automatic speech recognition, ASR)任务上也展现出优势。代表性方法包括基于流注意力机制的ASR模型,它通过综合多个麦克风阵列中各个流的后端概率,动态调整每个流的权重,提高分布式场景中语音识别的性能。本质上流注意力机制依赖于DNN和HMM的原理,在不增加计算复杂度的情况下,通过组合最可靠的声学特征提高流式ASR的准确性[168]。Wang等[169]提出了一种新型注意力框架,通过预测各个流的声学模型概率来增强性能,从而生成了更加精确的流注意力向量,并有效地整合了流所贡献的有用信息,该框架不仅提高了识别准确性,还提升了声场景鲁棒性和适应性。

为了降低分布式麦克风阵列ASR系统复杂度,文献[170]提出PickNet模型实现实时通道选择,如图9所示,该模型假设在每个时间点最多只有一个说话人处于活跃状态,并通过对短时间内的频谱片段进行识别以得到最接近活跃说话者的设备,并将选定麦克风的短时间信号帧串联起来生成输出信号。PickNet仅利用每个时间帧周围的有限声学上下文,因此能够实时运行,并且对声学条件的变化具有鲁棒性。即使增加输入通道数量,PickNet的计算成本也仅呈线性增长,因此适合实际分布式麦克风阵列应用场景中的低计算成本要求。

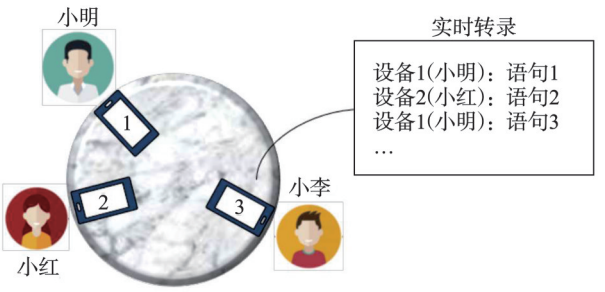

图9　分布式麦克风阵列场景多设备会话实时转录

Fig.9　Multi-devices real-time meeting transcription in the DMA scenarios

为了提高分布式阵列场景中ASR系统识别准确性,可以在识别模块前端使用拾音质量增强网络。相较于传统单麦克风阵列,分布式麦克风阵列具有任意摆放、覆盖范围面积大等优势,极有可能存在某个节点距离声源很近,准确选择最佳拾音阵列对提升语音质量以及可懂度至关重要。文献[171]在基于Transformer端到端ASR模型基础之上,通过对分布式麦克风阵列的余弦IPD特征和能量特征进行拼接,提出了分布式麦克风阵列的最佳拾音判定网络,判定阵列的拾音质量需要从目标声源和噪声源实际位置进行分析,从而确保复杂噪声条件下ASR性能。针对多说话人ASR任务,Yu等[172]面向多方会议场景利用帧级和信道级跨信道注意力之间的互补性,并同时建模信道和帧级信息,提出了多帧跨信道注意力方法。值得说明的是,在多说话人和大量随机分布麦克风情况下,可以使用说话人嵌入(embedding)或MFCC特征对麦克风进行聚类,每一类可以重点关注某说话人识别和ASR[173]。

汽车驾驶舱是分布式麦克风阵列ASR最常见的应用场景之一,与会话/会议场景的多人交谈ASR不同,车载ASR面临更多挑战,包括驾驶舱内复杂的声学环境,封闭和非常规的空间里有特殊的房间脉冲响应,导致了特殊的混响条件;舱内舱外存在风声、引擎声、轮胎声、背景音乐、说话干扰者等多种不同的噪声;不同的驾驶情况也会影响ASR表现,例如停车、高速、低速驾车、白天和夜晚驾驶等;缺乏大规模的公共真实车内数据。2023年年底,希尔贝壳、理想汽车、西北工业大学、新加坡南洋理工大学、天津大学、WeNet开源社区、微软、中国信通院等单位共同在IEEE国际声学、语音和信号处理会议(ICASSP2024)上发起了车载多通道语音识别挑战赛(In-Car Multi-Channel ASR, ICMC-ASR),赛事设置ASR和ASDR(Automatic speech diarization and recognition)两个赛道[174]。此次比赛构建起了1 000+小时车内真实录制的多通道、多说话人普通话语音数据,来源于车内不同座位的说话人,车内分布式麦克风与人员头戴麦克风分别收集了远场和近场数据,如图10所示。中科大-讯飞联合团队运用基于多音区声源定位的通道挑选、基于自监督表征学习和多说话人声纹特征的说话人角色分离、基

于多粒度单元增强的口音ASR等技术，最终分别以13.16%和21.48%的词错误率拿下两个赛道的第一名，与官方基线相比分别取得49.84%和70.52%的相对下降[175]。

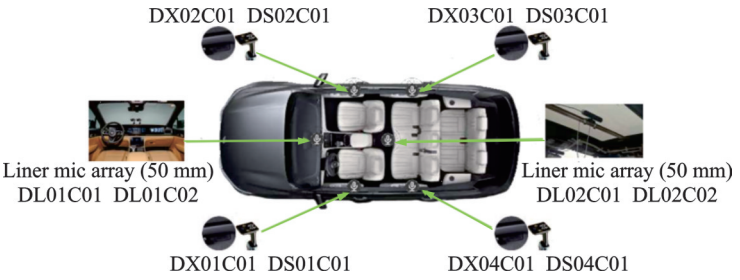

图10 基于分布式麦克风阵列的ICMC-ASR系统

Fig.10 DMA-based ICMC-ASR system

说话人日志(Speaker diarization)是智能会议系统的重要任务之一，要求将每个说话者的发言归属到相应的说话者。相比传统麦克风阵列系统，基于分布式麦克风阵列的说话人日志方法在应对重叠语音和远场录音时也展现出了显著的优势，因为很可能存在一个或多个距离发言者较近的麦克风设备。因此，研究者提出一种基于分布式麦克风的多通道端到端说话人日志方法[176]，该方法不依赖于麦克风的数量和通道数，能够处理不同数量和排列的麦克风输入，通过自注意力机制捕捉跨通道和跨帧的信息，从而实现对说话人变化的高效识别。其中，时空编码器和协同注意力编码器替换了传统Transformer编码器处理多通道输入，两者都独立于麦克风的数量和拓扑，意味着该方法具有良好的阵列阵型鲁棒性。另外，说话人日志也是ICMC-ASR挑战赛任务之一[174-175]，表明基于分布式麦克风阵列的说话人日志方法在车载语音交互场景也具有重要的应用潜力。

3.5 基于分布式麦克风阵列的音频场景分析

在音频场景分类任务中，通过利用多个空间分布的麦克风，能够捕捉更丰富的声学信息，从而提高音频场景分类的精度和鲁棒性。多通道卷积神经网络(Convolutional neural network, CNN)是处理分布式麦克风阵列信号的一种有效方法，能够同时处理多个麦克风捕捉的音频信号，通过卷积操作提取空间和时间特征。与单通道CNN相比，多通道CNN能够更好地捕捉空间信息，从而提升音频场景分类的性能。文献[177]利用多个空间分布的麦克风信号，通过卷积层提取各个通道的特征并融合，能够有效地利用空间信息、提升分类精度。更进一步，时空卷积网络结合了时间和空间卷积操作，能够同时捕捉音频信号的时间和空间特征。时空卷积网络通过在时间轴和空间轴上进行卷积，即对每个麦克风的时间信号进行卷积操作以提取时间特征，然后在空间轴上进行卷积，融合不同麦克风的空间特征，通过充分利用时间和空间特征，从而提升音频场景分类性能。

声音事件检测(Sound activity detection, SAD)是音频场景分析的核心任务之一(参阅DCASE比赛[178])，CNN也可用于分布式场景声音事件检测。文献[179]通过结构优化大幅减少了CNN的内存需求和计算复杂度，使其适用于分布式麦克风阵列环境，模型可以在每个传感器节点上高效运行，从而减少传输的数据量，通过在边缘节点进行SAD，可以显著降低传输需求和中央处理单元的负载，这对于分布式麦克风阵列的能耗管理尤为关键，并且该方法能够在每个节点上实现实时SAD，确保系统能够及时捕捉和处理重要事件。为了进一步提高SAD鲁棒性，可以使用自适应特征选择与融合方法，结合环境噪声和信号质量的变化动态调整每个麦克风的权重，从而在复杂环境中仍能保持较高的检测和分类性能，通过DNN自动学习和调整特征选择策略，能够在不同噪声条件下有效检测声音事件。另外，波束形成技术通过对多个麦克风信号进行时间对齐和加权求和，能够形成指向特定声源的波束，从而增强目标声音信号，抑制噪声和混响，前端结合该技术可以显著提升SAD与分类的准确性。

4　总结与展望

本文系统阐述了分布式麦克风阵列拾音层面的组织理论、基于节点效用评估的阵列布局优化理论,以及面向典型语音交互任务的应用方法。充分表明,分布式麦克风阵列在智能家居、会议、声场景监控、智慧城市、助听器、车载语音交互等场景展现出重要的应用潜力,其灵活的布阵方式和广阔的声场覆盖度有益于改善语音增强、声源定位、语音识别、说话人日志和音频场景分析等任务性能,基本实现了对传统规则型麦克风阵列所能处理的语音任务的全覆盖,并且有望构建更加智能的全空间均匀式拾音系统。需要强调的是不同于传统规则型麦克风阵列,波束形成、时延补偿等环节并不是分布式麦克风阵列所关注的重点,布阵效率和拾音质量才是其核心问题,波束形成是前文所述的分布式麦克风阵列语音增强方法的关键技术之一。另外,本文未针对性地介绍盲源分离、声源数量确定、噪声/回声抑制和去混响等其他任务,因为这些任务基本都可以运用语音增强、声源定位等技术予以解决。

分布式麦克风阵列虽然为新一代拾音方案提供了更多机遇,但是相较于传统麦克风阵列也带来了一些新的挑战性问题,包括时变未知的拓扑结构、分布式数据融合、有限通信带宽和能量资源、阵型灵活拓展、设备时钟同步、拾音-反馈延时、通信层与物理层之间跨层设计等等[7],前面章节对部分问题已有应对方法,但这些方法通常为独立设计,缺乏有机集成考量。例如,依据2022年北京冬奥会雪上项目的特点及相关制作规范,雪上技巧、空中技巧、U型池和平行大回转4个雪上场地赛场内的拾音设计方案表明,真实场景中麦克风布局设计与拾音质量需要综合多方面因素平衡考虑[180];实际应用基于(相对)声学传递函数的分布式语音增强和基于TOA/TDOA/DOA的声源定位方法需要配合麦克风节点自定位方法[14-17],从而提供麦克风位置信息,而时钟同步方法又是所有分布式语音处理方法的必要前提。因此,阵列组织理论与拾音应用需求之间的有机融合和多模块协同优化是分布式麦克风阵列走向实用的重要挑战之一。

其次,现阶段分布式麦克风阵列在智能会议(腾讯天籁inside音频解决方案[9])、智能家居(科大讯飞MORFEI家居物联平台[181])等场景已有落地应用,但它们本质上还属于集中式阵列管理模式,尚未考虑无线局域网通信模式下的数据编码、路由、传输与分布式计算。分布式数据处理方式(去中心化)应该是分布式麦克风阵列的标志,这也是分布式麦克风阵列的重要研究方向之一。

再次,隐私保护是分布式麦克风阵列设计必须考虑的问题之一[21,182-184]。理论上声场景所有的无线设备均可以参与构建分布式麦克风阵列,但是在开放环境中用户会担心自身手机、电脑、iPad等终端会遭到恶意攻击、与目标语音感知任务无关的数据发生泄漏,特别是在当今深度学习技术快速进步和攻击手段日益隐蔽化复杂化的时代。隐私泄漏在相对封闭的智能家居场景下或许不是问题,但为了加速分布式麦克风阵列走向更多实际场景的应用,限制节点访问无关数据、解决隐私保护问题是构建分布式麦克风阵列拾音系统的重要条件。

最后,现阶段分布式麦克风阵列所处理的任务比较单一,并且存在严重的信息冗余问题,不能满足实际语音交互场景下多任务处理需求[12]。理想的情况应该是结合任务需求针对大规模无线设备进行聚类[173,185-187],一类主要负责一项任务,类与类之间应建立协同分工关系,从而降低数据量和系统复杂度。麦克风聚类也可以一定程度上缓解隐私泄漏,其中聚类准则、复合信息解耦和类间协同分工机制是设计多任务型分布式麦克风阵列拾音系统的重要研究方向。

总之,分布式麦克风阵列是随着智能终端设备的大量普及而涌现的新型拾音与语音交互系统,经过过去十多年的发展分布式麦克风阵列技术取得了巨大进步,但是离大规模应用还存在不小差距。未来需要充分挖掘分布式麦克风阵列的无线传感器网络和麦克风阵列双重属性,运用传感器网络、语音信号处理、阵列信号处理、信息安全和人工智能等多学科交叉融合的方式,实现“阵列组织、布局优化、拾音应用”的一体化设计,从而满足分布式麦克风阵列的实用需求。

参考文献：

[1] BRANDSTEIN M, DARREN W. Microphone arrays: Signal processing techniques and applications[M]. New York: Springer, 2001.

[2] RYAN J G, GOUBRAN R A. Array optimization applied in the near field of a microphone array[J]. IEEE Transactions on Speech and Audio Processing, 2000, 8(2): 173-176.

[3] BENESTY J, CHEN J, HUANG Y, et al. On microphone-array beamforming from a MIMO acoustic signal processing perspective[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 1053-1065.

[4] LEE H. Multichannel 3D microphone arrays: A review[J]. Journal of the Audio Engineering Society, 2021, 69(1/2): 5-26.

[5] PÖNTYNEN H, SALMINEN N H. Resolving front-back ambiguity with head rotation: The role of level dynamics[J]. Hearing Research, 2019, 377: 196-207.

[6] AYLLÓN D, GIL-PITA R, ROSA-ZURERA M. Design of microphone arrays for hearing aids optimized to unknown subjects [J]. Signal Processing, 2013, 93(11): 3239-3250.

[7] BERTRAND A. Applications and trends in wireless acoustic sensor networks: A signal processing perspective[C]// Proceedings of 2011 18th IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT). Ghent: IEEE, 2011: 1-6.

[8] ALÍAS F, ALSINA-PAGÈS R M. Review of wireless acoustic sensor networks for environmental noise monitoring in smart cities[J]. Journal of Sensors, 2019, 2019: 7634860.

[9] 腾讯天籁 . 专治会议室“听不清”,腾讯天籁 inside 音频解决方案升级版[EB/OL].(2022-08-26). https://meeting.tencent.com/news/inside-Upgraded-version.html.

[10] ZHANG J, LI C. Quantization-aware binaural MWF based noise reduction incorporating external wireless devices[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3118-3131.

[11] ZHANG J, HEUSDENS R, HENDRIKS R C. Rate-distributed binaural LCMV beamforming for assistive hearing in wireless acoustic sensor networks[C]//Proceedings of 2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM). Sheffield: IEEE, 2018: 460-464.

[12] HASSANI A, PLATA-CHAVES J, BAHARI M H, et al. Multi-task wireless sensor network for joint distributed node-specific signal enhancement, LCMV beamforming and DOA estimation[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(3): 518-533.

[13] TURCHET L, FAZEKAS G, LAGRANGE M, et al. The internet of audio things: State of the art, vision, and challenges[J]. IEEE Internet of Things Journal, 2020, 7(10): 10233-10249.

[14] GAUBITCH N D, KLEIJN W B, HEUSDENS R. Auto-localization in ad-hoc microphone arrays[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013: 106-110.

[15] WANG L, HON T K, REISS J D, et al. Self-localization of ad-hoc arrays using time difference of arrivals[J]. IEEE Transactions on Signal Processing, 2016, 64(4): 1018-1033.

[16] HON T K, WANG L, REISS J D, et al. Audio fingerprinting for multi-device self-localization[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(10): 1623-1636.

[17] WANG X, HU D. Distributed self-localization for acoustic transceiver networks[J]. IEEE Signal Processing Letters, 2023, 30: 553-557.

[18] SHAH S, BEFERULL-LOZANO B. Adaptive quantization for multihop progressive estimation in wireless sensor networks [C]//Proceedings of 21st European Signal Processing Conference (EUSIPCO 2013). Marrakech: IEEE, 2013: 1-5.

[19] HUANG Y, HUA Y. Multihop progressive decentralized estimation in wireless sensor networks[J]. IEEE Signal Processing Letters, 2007, 14(12): 1004-1007.

[20] HUANG Y, HUA Y. Energy planning for progressive estimation in multihop sensor networks[J]. IEEE Transactions on Signal Processing, 2009, 57(10): 4052-4065.

[21] HENDRIKS R C, ERKIN Z, GERKMANN T. Privacy-preserving distributed speech enhancement forwireless sensor

networks by processing in the encrypted domain[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013: 7005-7009.

[22] FRAMPTON K D, BAUMANN O N, GARDONIO P. A comparison of decentralized, distributed, and centralized vibro-acoustic control[J]. The Journal of the Acoustical Society of America, 2010, 128(5): 2798-2806.

[23] DOCLO S, MOONEN M. GSVD-based optimal filtering for single and multimicrophone speech enhancement[J]. IEEE Transactions on Signal Processing, 2002, 50(9): 2230-2244.

[24] MARKOVICH S, GANNOT S, COHEN I. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009, 17(6): 1071-1086.

[25] ZENG Y, HENDRIKS R C. Distributed delay and sum beamformer for speech enhancement *via* randomized gossip[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(1): 260-273.

[26] JIA Y, LUO Y, LIN Y, et al. Distributed microphone arrays for digital home and office[C]//Proceedings of 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Toulouse: IEEE, 2006.

[27] HIMAWAN I, MCCOWAN I, SRIDHARAN S. Clustered blind beamforming from ad-hoc microphone arrays[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 661-676.

[28] HIOKA Y, KLEIJN W B. Distributed blind source separation with an application to audio signals[C]//Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Prague: IEEE, 2011: 233-236.

[29] SHANNON C E. Communication in the presence of noise[C]//Proceedings of the IRE. [S.l]:IEEE, 1949: 10-21.

[30] ROY O, VETTERLI M. Rate-constrained collaborative noise reduction for wireless hearing aids[J]. IEEE Transactions on Signal Processing, 2009, 57(2): 645-657.

[31] AMINI J, HENDRIKS R C, HEUSDENS R, et al. Spatially correct rate-constrained noise reduction for binaural hearing aids in wireless acoustic sensor networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2731-2742.

[32] AMINI J, HENDRIKS R C, HEUSDENS R, et al. Asymmetric coding for rate-constrained noise reduction in binaural hearing aids[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(1): 154-167.

[33] BERTRAND A, MOONEN M. Distributed adaptive node-specific signal estimation in fully connected sensor networks: Part I: Sequential node updating[J]. IEEE Transactions on Signal Processing, 2010, 58(10): 5277-5291.

[34] BERTRAND A, MOONEN M. Distributed adaptive node-specific signal estimation in fully connected sensor networks: Part I: Sequential node updating[J]. IEEE Transactions on Signal Processing, 2010, 58(10): 5277-5291.

[35] BERTRAND A, MOONEN M. Distributed adaptive estimation of node-specific signals in wireless sensor networks with a tree topology[J]. IEEE Transactions on Signal Processing, 2011, 59(5): 2196-2210.

[36] HASSANI A, BERTRAND A, MOONEN M. GEVD-based low-rank approximation for distributed adaptive node-specific signal estimation in wireless sensor networks[J]. IEEE Transactions on Signal Processing, 2016, 64(10): 2557-2572.

[37] CHERKASSKY D, GANNOT S. Blind synchronization in wireless sensor networks with application to speech enhancement [C]//Proceedings of 2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC). Juan-les-Pins: IEEE, 2014: 183-187.

[38] SCHMALENSTROEER J, HAEB-UMBACH R. Sampling rate synchronisation in acoustic sensor networks with a pre-trained clock skew error model[C]//Proceedings of 21st European Signal Processing Conference (EUSIPCO 2013). Marrakech: IEEE, 2013: 1-5.

[39] SCHMALENSTROEER J, JEBRAMCIK P, HAEB-UMBACH R. A gossiping approach to sampling clock synchronization in wireless acoustic sensor networks[C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 7575-7579.

[40] MARKOVICH-GOLAN S, GANNOT S, COHEN I. Blind sampling rate offset estimation and compensation in wireless acoustic sensor networks with application to beamforming[C]//Proceedings of IWAENC 2012; International Workshop on

Acoustic Signal Enhancement. Aachen: VDE, 2012: 1-4.

[41] SCHMALENSTROEER J, HEYMANN J, DRUDE L, et al. Multi-stage coherence drift based sampling rate synchronization for acoustic beamforming[C]//Proceedings of 2017 IEEE 19th International Workshop on Multimedia Signal Processing (MMSP). Luton: IEEE, 2017: 1-6.

[42] ZHANG J, WU P. Joint sampling synchronization and source localization for wireless acoustic sensor networks[J]. IEEE Communications Letters, 2020, 24(5): 1020-1023.

[43] HU D, ZHANG H, BAO F, et al. Distributed sampling rate offset estimation over acoustic sensor networks based on asynchronous network newton optimization[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1952, 31: 301-312.

[44] CHINAEV A, ENZNER G, GBURREK T, et al. Online estimation of sampling ate offsets in wireless acoustic sensor networks with packet loss[C]//Proceedings of EUSIPCO.[S.l.]:IEEE, 2021: 1110-1114.

[45] CHERKASSKY D, MARKOVICH-GOLAN S, GANNOT S. Performance analysis of MVDR beamformer in WASN with sampling rate offsets and blind synchronization[C]//Proceedings of 2015 23rd European Signal Processing Conference (EUSIPCO). Nice: IEEE, 2015: 245-249.

[46] WANG L, DOCLO S. Correlation maximization-based sampling rate offset estimation for distributed microphone arrays[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 571-582.

[47] MIYABE S, ONO N, MAKINO S. Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver: IEEE, 2013: 674-678.

[48] MIYABE S, ONO N, MAKINO S. Optimizing frame analysis with non-integer shift for sampling mismatch compensation of long recording[C]//Proceedings of WASPAA. [S.l.]: IEEE, 2013: 1-4.

[49] ARAKI S, ONO N, KINOSHITA K, et al. Estimation of sampling frequency mismatch between distributed asynchronous microphones under existence of source movements with stationary time periods detection[C]//Proceedings of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 785-789.

[50] MASUYAMA Y, YAMAOKA K, KAWAMURA T, et al. Efficient joint optimization of sampling rate offsets using entire multichannel signal[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1816, 32: 1816-1828.

[51] CHINAEV A, THÜNE P, ENZNER G. A double-cross-correlation processor for blind sampling rate offset estimation in acoustic sensor networks[C]//Proceedings of ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 641-645.

[52] CHINAEV A, ENZNER G. Distributed synchronization for ad-hoc acoustic sensor networks using closed-loop double-cross-correlation processing[C]//Proceedings of 2022 International Workshop on Acoustic Signal Enhancement (IWAENC). Bamberg: IEEE, 2022: 1-5.

[53] CHINAEV A, KNAEPPER N, ENZNER G. Online distributed waveform-synchronization for acoustic sensor networks with dynamic topology[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 55.

[54] CHINAEV A, THÜNE P, ENZNER G. Double-cross-correlation processing for blind sampling-rate and time-offset estimation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 1881, 29: 1881-1896.

[55] CHINAEV A, KNAEPPER N, ENZNER G. Long-term synchronization of wireless acoustic sensor networks with nonpersistent acoustic activity using coherence state[C]//Proceedings of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island: IEEE, 2023: 1-5.

[56] GUNTHER M, AFIFI H, BRENDEL A, et al. Network-aware optimal microphone channel selection in wireless acoustic sensor networks[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 820-824.

[57] GÜNTHER M, BRENDEL A, KELLERMANN W. Online estimation of time-variant microphone utility in wireless acoustic sensor networks using single-channel signal features[C]//Proceedings of 2021 29th European Signal Processing Conference

(EUSIPCO). Dublin: IEEE, 2021: 1120-1124.

[58] GÜNTHER M, BRENDEL A, KELLERMANN W. Microphone utility estimation in acoustic sensor networks using single-channel signal features[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 1-19.

[59] NARAYANAN A M, BERTRAND A. Group-utility metric for efficient sensor selection and removal in LCMV beamformers [C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona : IEEE, 2020: 4950-4954.

[60] BERTRAND A. Utility metrics for assessment and subset selection of input variables for linear estimation[tips & tricks][J]. IEEE Signal Processing Magazine, 2018, 35(6): 93-99.

[61] SZURLEY J, BERTRAND A, MOONEN M. Efficient computation of microphone utility in a wireless acoustic sensor network with multi-channel Wiener filter based noise reduction[C]//Proceedings of ICASSP.[S.l.]: IEEE, 2012: 2657-2660.

[62] DOCLO S, SPRIET A, WOUTERS J, et al. Speech distortion weighted multichannel Wiener filtering techniques for noise reduction[M]//Signals and Communication Technology. Heidelberg: Springer-Verlag, 2005: 199-228.

[63] ZHANG J, TAO R, DU J, et al. SDW-SWF: Speech distortion weighted single-channel Wiener filter for noise reduction[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 3176-3189.

[64] SZURLEY J, BERTRAND A, MOONEN M, et al. Energy aware greedy subset selection for speech enhancement in wireless acoustic sensor networks[C]//Proceedings of 2012 the 20th European Signal Processing Conference (EUSIPCO). Bucharest: IEEE, 2012: 789-793.

[65] ZHANG J, CHEPURI S P, HENDRIKS R C, et al. Microphone subset selection for MVDR beamformer based noise reduction[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(3): 550-563.

[66] ZHANG J, DU J, DAI L R. Sensor selection for relative acoustic transfer function steered linearly-constrained beamformers[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1220-1232.

[67] ZHANG J, ZHANG G, DAI L. Frequency-invariant sensor selection for MVDR beamforming in wireless acoustic sensor networks[J]. IEEE Transactions on Wireless Communications, 2022, 21(12): 10648-10661.

[68] ZHANG J, HEUSDENS R, HENDRIKS R C. Relative acoustic transfer function estimation in wireless acoustic sensor networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(10): 1507-1519.

[69] MARKOVICH-GOLAN S, GANNOT S. Performance analysis of the covariance subtraction method for relative transfer function estimation and comparison to the covariance whitening method[C]//Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane: IEEE, 2015: 544-548.

[70] MARKOVICH-GOLAN S, GANNOT S, KELLERMANN W. Performance analysis of the covariance-whitening and the covariance-subtraction methods for estimating the relative transfer function[C]//Proceedings of 2018 26th European Signal Processing Conference (EUSIPCO). Rome: IEEE, 2018: 2499-2503.

[71] ZHANG J, CHEN H, DAI L R, et al. A study on reference microphone selection for multi-microphone speech enhancement [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 671-683.

[72] ZHANG J, HEUSDENS R, HENDRIKS R C. Rate-distributed spatial filtering based noise reduction in wireless acoustic sensor networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(11): 2015-2026.

[73] AMAR A, DORON M A. A linearly constrained minimum variance beamformer with a pre-specified suppression level over a pre-defined broad null sector[J]. Signal Processing, 2015, 109: 165-171.

[74] ZHANG J, HEUSDENS R, HENDRIKS R C. Sensor selection and rate distribution based beamforming for wireless acoustic sensor networks[C]//Proceedings of EUSIPCO. [S.l.]: IEEE, 2019: 1-5.

[75] ZHANG J, KOUTROUVELIS A I, HEUSDENS R, et al. Distributed rate-constrained LCMV beamforming[J]. IEEE Signal Processing Letters, 2019, 26(5): 675-679.

[76] ZHANG G, HEUSDENS R. Distributed optimization using the primal-dual method of multipliers[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2018, 4(1): 173-187.

[77] ZHANG J, TAO R, DU J, et al. Energy-efficient sparsity-driven speech enhancement in wireless acoustic sensor networks[J].

IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 215-228.

[78] HU D, WANG X, LIU R, et al. Fast subnetwork selection for speech enhancement in wireless acoustic sensor networks[C]//Proceedings of 2023 8th International Conference on Signal and Image Processing (ICSIP). Wuxi: IEEE, 2023: 900-904.

[79] HU D, SI Q, LIU R, et al. Distributed sensor selection for speech enhancement with acoustic sensor networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 985-999.

[80] HU D, SI Q, BAO F, et al. Distributed energy-saving speech enhancement in wireless acoustic sensor networks[J]. Information Fusion, 2025, 113: 102593.

[81] TAVAKOLI V M, JENSEN J R, CHRISTENSEN M G, et al. A framework for speech enhancement with ad hoc microphone arrays[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(6): 1038-1051.

[82] TAVAKOLI V M, JENSEN J R, HEUSDENS R, et al. Distributed max-SINR speech enhancement with ad hoc microphone arrays[C]//Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017: 151-155.

[83] BOYD S. Distributed optimization and statistical learning *via* the alternating direction method of multipliers[J]. Foundations and Trends® in Machine Learning, 2010, 3(1): 1-122.

[84] BOYD S, GHOSH A, PRABHAKAR B, et al. Randomized gossip algorithms[J]. IEEE Transactions on Information. Theory, 2004, 52(6): 2508-2530.

[85] YANG T, YI X, WU J, et al. A survey of distributed optimization[J]. Annual Reviews in Control, 2019, 47: 278-305.

[86] ZENG Y, HENDRIKS R C. Distributed delay and sum beamformer for speech enhancement in wireless sensor networks *via* randomized gossip[C]//Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Kyoto: IEEE, 2012: 4037-4040.

[87] BERTRAND A, CALLEBAUT J, MOONEN M. Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks[C]//Proceedings of IWAENC. [S.l.]: IEEE, 2010: 2-5.

[88] BERTRAND A, MOONEN M. Distributed LCMV beamforming in a wireless sensor network with single-channel per-node signal transmission[J]. IEEE Transactions on Signal Processing, 2013, 61(13): 3447-3459.

[89] MARKOVICH-GOLAN S, GANNOT S, COHEN I. Distributed multiple constraints generalized sidelobe canceler for fully connected wireless acoustic sensor networks[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(2): 343-356.

[90] SZURLEY J, BERTRAND A, MOONEN M. Distributed adaptive node-specific signal estimation in heterogeneous and mixed-topology wireless sensor networks[J]. Signal Processing, 2015, 117: 44-60.

[91] MARKOVICH-GOLAN S, BERTRAND A, MOONEN M, et al. Optimal distributed minimum-variance beamforming approaches for speech enhancement in wireless acoustic sensor networks[J]. Signal Processing, 2015, 107: 4-20.

[92] SZURLEY J, BERTRAND A, MOONEN M. Topology-independent distributed adaptive node-specific signal estimation in wireless sensor networks[J]. IEEE Transactions on Signal and Information Processing Over Networks, 2017, 3(1): 130-144.

[93] HEUSDENS R, ZHANG G, HENDRIKS R C, et al. Distributed MVDR beamforming for (wireless) microphone networks using message passing[C]//Proceedings of IWAENC 2012; International Workshop on Acoustic Signal Enhancement. Aachen: VDE, 2012: 1-4.

[94] O′CONNOR M, KLEIJN W B. Diffusion-based distributed MVDR beamformer[C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence: IEEE, 2014: 810-814.

[95] SHERSON T, KLEIJN W B, HEUSDENS R. A distributed algorithm for robust LCMV beamforming[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 101-105.

[96] O′CONNOR M, KLEIJN W B, ABHAYAPALA T. Distributed sparse MVDR beamforming using the bi-alternating direction method of multipliers[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 106-110.

[97] CHANG R, CHEN Z, YIN F. Distributed multichannel Wiener filtering for speech enhancement in acoustic sensor networks [J]. International Journal of Adaptive Control and Signal Processing, 2022, 36(11): 2732-2753.

[98] FURNON N, SERIZEL R, ESSID S, et al. DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2838, 29: 2310-2323.

[99] ZHANG X L. Deep ad-hoc beamforming[J]. Computer & Speech Language, 2021, 68: 101201.

[100] ZHANG S Y, LI X F. Microphone array generalization for multichannel narrowband deep speech enhancement[EB/OL].(2021-07-27). Http://doi.org/10.48550/arXiv.2107.12601.

[101] NEO V W, EVERS C, NAYLOR P A. Enhancement of noisy reverberant speech using polynomial matrix eigenvalue decomposition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3255-3266.

[102] JENSEN J R, BENESTY J, CHRISTENSEN M G. Noise reduction with optimal variable span linear filters[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(4): 631-644.

[103] D'OLNE E, NEO V W, NAYLOR P A. Speech enhancement in distributed microphone arrays using polynomial eigenvalue decomposition[C]//Proceedings of 2022 30th European Signal Processing Conference (EUSIPCO). Belgrade: IEEE, 2022: 55-59.

[104] BERTRAND A, MOONEN M. Distributed adaptive generalized eigenvector estimation of a sensor signal covariance matrix pair in a fully connected sensor network[J]. Signal Processing, 2015, 106: 209-214.

[105] HASSANI A, BERTRAND A, MOONEN M. LCMV beamforming with subspace projection for multi-speaker speech enhancement[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai: IEEE, 2016: 91-95.

[106] LIU H, DARABI H, BANERJEE P, et al. Survey of wireless indoor positioning techniques and systems[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2007, 37(6): 1067-1080.

[107] MANOLAKIS D E. Efficient solution and performance analysis of 3-D position estimation by trilateration[J]. IEEE Transactions on Aerospace and Electronic Systems, 1996, 32(4): 1239-1248.

[108] ZHOU Y. An efficient least-squares trilateration algorithm for mobile robot localization[C]//Proceedings of 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems. St. Louis: IEEE, 2009: 3474-3479.

[109] ZOU Y, LIU H. A simple and efficient iterative method for ToA localization[C]//Proceedings of ICASSP. [S.l.]: IEEE, 2020: 4881-4884.

[110] COBOS M, PEREZ-SOLANO J J, FELICI-CASTELL S, et al. Cumulative-sum-based localization of sound events in low-cost wireless acoustic sensor networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1792-1802.

[111] COBOS M, PEREZ-SOLANO J J, BELMONTE Ó, et al. Simultaneous ranging and self-positioning in unsynchronized wireless acoustic sensor networks[J]. IEEE Transactions on Signal Processing, 2016, 64(22): 5993-6004.

[112] ZHANG J, HENDRIKS R C, HEUSDENS R. Structured total least squares based internal delay estimation for distributed microphone auto-localization[C]//Proceedings of 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC). Xi'an: IEEE, 2016: 1-5.

[113] CHAN Y T, HO K C. A simple and efficient estimator for hyperbolic location[J]. IEEE Transactions on Signal Processing, 1994, 42(8): 1905-1915.

[114] JYOTHI R, BABU P. SOLVIT: A reference-free source localization technique using majorization minimization[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2661-2673.

[115] AJDLER T, KOZINTSEV I, LIENHART R, et al. Acoustic source localization in distributed sensor networks[C]//Proceedings of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers. Pacific: IEEE, 2004: 1328-1332.

[116] COMPAGNONI M, BESTAGINI P, ANTONACCI F, et al. Localization of acoustic sources through the fitting of propagation cones using multiple independent arrays[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(7): 1964-1975.

[117] JIA J, LIU M, LI X. Acoustic passive localization algorithm based on wireless sensor networks[C]//Proceedings of 2009 International Conference on Mechatronics and Automation. Changchun: IEEE, 2009: 1145-1149.

[118] JIA J, LIU M, LI X. Acoustic localization algorithm using wireless sensor networks[C]//Proceedings of 2009 the Second International Conference on Intelligent Computation Technology and Automation. Changsha: IEEE, 2009: 434-437.

[119] CANCLINI A, BESTAGINI P, ANTONACCI F, et al. A robust and low-complexity source localization algorithm for asynchronous distributed microphone networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(10): 1563-1575.

[120] DANG X, MA W, HABETS E A P, et al. TDOA-based robust sound source localization with sparse regularization in wireless acoustic sensor networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2956, 30: 1108-1123.

[121] WANG Y, HO K C, WANG Z. Robust localization under NLOS environment in the presence of isolated outliers by full-Set TDOA measurements[J]. Signal Processing, 2023, 212: 109159.

[122] COMPAGNONI M, PINI A, CANCLINI A, et al. A geometrical-statistical approach to outlier removal for TDOA measurements[J]. IEEE Transactions on Signal Processing, 2017, 65(15): 3960-3975.

[123] LU Z. Sound event detection and localization based on CNN and LSTM[R]. [S.l.]: [s.n.], 2019.

[124] NOH K, CHOI J, JEON D, et al. Three-stage approach for sound event localization and detection [R]. [S.l.]: [s.n.], 2019.

[125] VERA-DIAZ J M, PIZARRO D, MACIAS-GUARASA J. Towards domain independence in CNN-based acoustic localization using deep cross correlations[C]//Proceedings of 2020 28th European Signal Processing Conference (EUSIPCO). Amsterdam: IEEE, 2021: 226-230.

[126] YU W, GAUBITCH N D, HEUSDENS R. Distributed TDOA-based indoor source localization[C]//Proceedings of ICASSP. [S.l.]: IEEE, 2018: 6887-6891.

[127] GENDLER A, YARON PELEG S, AMAR A. A diffusion-based distributed time difference of arrival source positioning[C]//Proceedings of 2021 IEEE 24th International Conference on Information Fusion (FUSION). Sun City: IEEE, 2021: 1-5.

[128] MEESOOKHO C, MITRA U, NARAYANAN S. On energy-based acoustic source localization for sensor networks[J]. IEEE Transactions on Signal Processing, 2008, 56(1): 365-377.

[129] WANG G, CHEN H, LI Y, et al. On received-signal-strength based localization with unknown transmit power and path loss exponent[J]. IEEE Wireless Communications Letters, 2012, 1(5): 536-539.

[130] GHOLAMI M R, VAGHEFI R M, STRÖM E G. RSS-based sensor localization in the presence of unknown channel parameters[J]. IEEE Transactions on Signal Processing, 2013, 61(15): 3752-3759.

[131] VAGHEFI R M, GHOLAMI M R, BUEHRER R M, et al. Cooperative received signal strength-based sensor localization with unknown transmit powers[J]. IEEE Transactions on Signal Processing, 2013, 61(6): 1389-1403.

[132] HU Y, LEUS G. Robust differential received signal strength-based localization[J]. IEEE Transactions on Signal Processing, 2017, 65(12): 3261-3276.

[133] 张祺, 巩朋成, 郑毅豪, 等. 基于模式空间算法的声源二维DOA估计[J]. 数据采集与处理, 2020, 35(5): 867-879.
ZHANG Qi, GONG Pengcheng, ZHENG Yihao, et al. Two-dimensional DOA estimation of sound source based on mode space algorithm[J]. Journal of Data Acquisition and Processing, 2020, 35(5): 867-879.

[134] ROY R, KAILATH T. ESPRIT-estimation of signal parameters *via* rotational invariance techniques[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37(7): 984-995.

[135] GRONDIN F, GLASS J. SVD-PHAT: A fast sound source localization method[C]//Proceedings of ICASSP. [S.l.]: IEEE, 2019: 4140-4144.

[136] STANSFIELD R G. Statistical theory of d. f. fixing[J]. Journal of the Institution of Electrical Engineers-Part IIIA: Radiocommunication, 1947, 94(15): 762-770.

[137] TORRIERI D J. Statistical theory of passive location systems[J]. IEEE Transactions on Aerospace and Electronic Systems, 1984, 20(2): 183-198.

[138] DOĞANÇAY K. Bearings-only target localization using total least squares[J]. Signal Processing, 2005, 85(9): 1695-1710.

[139] WANG Z, LUO J A, ZHANG X P. A novel location-penalized maximum likelihood estimator for bearing-only target localization[J]. IEEE Transactions on Signal Processing, 2012, 60(12): 6166-6181.

[140] WANG Y, HO K C. An asymptotically efficient estimator in closed-form for 3-D AOA localization using a sensor network[J]. IEEE Transactions on Wireless Communications, 2015, 14(12): 6524-6535.

[141] DIBIASE J H, SILVERMAN H F, BRANDSTEIN M S. Robust localization in reverberant rooms[M]//BRANDSTEIN M, WARD D, eds. Digital Signal Processing. Heidelberg: Springer Berlin Heidelberg, 2001: 157-180.

[142] ASTAPOV S, PREDEN J S, BERDNIKOVA J. Simplified acoustic localization by linear arrays for wireless sensor networks [C]//Proceedings of 2013 18th International Conference on Digital Signal Processing (DSP). Fira: IEEE, 2013: 1-6.

[143] ASTAPOV S, BERDNIKOVA J, PREDEN J S. Optimized acoustic localization with SRP-PHAT for monitoring in distributed sensor networks[J]. International Journal of Electronics and Telecommunications, 2013, 59(4): 383-390.

[144] AARABI P. The fusion of distributed microphone arrays for sound localization[J]. EURASIP Journal on Advances in Signal Processing, 2003, 2003(4): 860465.

[145] COBOS M, MARTI A, LOPEZ J J. A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling[J]. IEEE Signal Processing Letters, 2011, 18(1): 71-74.

[146] LIMA M V S, MARTINS W A, NUNES L O, et al. A volumetric SRP with refinement step for sound source localization[J]. IEEE Signal Processing Letters, 2015, 22(8): 1098-1102.

[147] SALVATI D, DRIOLI C, FORESTI G L. Acoustic source localization using a geometrically sampled grid SRP-PHAT algorithm with max-pooling operation[J]. IEEE Signal Processing Letters, 2022, 29: 1828-1832.

[148] YIN J, VERHELST M. CNN-based robust sound source localization with SRP-PHAT for the extreme edge[J]. ACM Transactions on Embedded Computing Systems, 2023, 22(3): 1-27.

[149] LIU M, HU J, ZENG Q, et al. Sound source localization based on multi-channel cross-correlation weighted beamforming[J]. Micromachines, 2022, 13(7): 1010.

[150] ÇAKMAK B, DIETZEN T, ALI R, et al. A distributed steered response power approach to source localization in wireless acoustic sensor networks[C]//Proceedings of 2022 International Workshop on Acoustic Signal Enhancement (IWAENC). Bamberg: IEEE, 2022: 1-5.

[151] LAUFER-GOLDSHTEIN B, TALMON R, GANNOT S. Semi-supervised source localization on multiple manifolds with distributed microphones[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(7): 1477-1491.

[152] JING Y, LI Z, LIU C. Acoustic source tracking based on adaptive distributed particle filter in distributed microphone networks [J]. Signal Processing, 2019, 154: 375-386.

[153] 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.

[154] MCCOWAN I, PELECANOS J W, SRIDHARAN S. Robust speaker recognition using microphone arrays[C]//Proceedings of Odyssey. [S.l.]: IEEE, 2001: 101-106.

[155] XU R, MEI G, REN Z, et al. A real time speaker verification demonstration on the smart flow system[C]//Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing. Hong Kong, China: IEEE, 2004: 226-229.

[156] JAYANNA H S, MAHADEVA PRASANNA S R. Analysis, feature extraction, modeling and testing techniques for speaker recognition[J]. IETE Technical Review, 2009, 26(3): 181.

[157] GU R, ZHANG S X, CHEN L, et al. Enhancing end-to-end multi-channel speech separation *via* spatial feature learning[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 7319-7323.

[158] WANG Y, ZHANG J, CHEN S, et al. A study of multichannel spatiotemporal features and knowledge distillation on robust target speaker extraction[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Seoul: IEEE, 2024: 431-435.

[159] CAMPBELL J P. Speaker recognition: A tutorial[J]. Proceedings of the IEEE, 1997, 85(9): 1437-1462.

[160] REYNOLDS D A. An overview of automatic speaker recognition technology[C]//Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Orlando: IEEE, 2002: IV-4072-IV-4075.

[161] BAI Z, ZHANG X L. Speaker recognition based on deep learning: An overview[J]. Neural Networks, 2021, 140: 65-99.

[162] ANGUERA X, WOOTERS C, HERNANDO J. Acoustic beamforming for speaker diarization of meetings[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(7): 2011-2022.

[163] 孙磊. 复杂声学场景下多人对话语音识别的预处理方法研究[D]. 合肥: 中国科学技术大学, 2020.
SUN Lei. Research on preprocessing method of multi-person dialogue speech recognition in complex acoustic scene[D]. Hefei: University of Science and Technology of China, 2020.

[164] GUAN S Z, LIU S P, CHEN J Q, et al. Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays[C]//Proceedings of APSIPA ASC. [S.l.]: IEEE, 2021: 1116-1120.

[165] LIANG C, CHEN Y, YAO J, et al. Multi-channel far-field speaker verification with large-scale ad-hoc microphone arrays[C]//Proceedings of Interspeech 2022. [S.l.]: ISCA, 2022: 3679-3683.

[166] LIANG C, CHEN J, GUAN S, et al. Attention-based multi-channel speaker verification with ad-hoc microphone arrays[C]//Proceedings of 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Tokyo: IEEE, 2021: 1111-1115.

[167] LI G, YU J, DENG J, et al. Audio-visual multi-channel speech separation, dereverberation and recognition[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 6042-6046.

[168] KIM S, LANE I. Recurrent models for auditory attention in multi-microphone distant speech recognition[C]//Proceedings of Interspeech 2016. [S.l.]: ISCA, 2016: 3838-3842.

[169] WANG X, LI R, HERMANSKY H. Stream attention for distributed multi-microphone speech recognition[C]//Proceedings of Interspeech 2018. [S.l.]: ISCA, 2018: 3033-3037.

[170] YOSHIOKA T, WANG X, WANG D. PickNet: Real-time channel selection for ad hoc microphone arrays[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 921-925.

[171] 卢卓城. 基于分布式麦克风阵列的远场语音识别[D]. 深圳: 深圳大学, 2022.
LU Zhuocheng. Far-field speech recognition based on distributed microphone array[D]. Shenzhen: Shenzhen University, 2022.

[172] YU F, ZHANG S, GUO P, et al. MFCCA: Multi-frame cross-channel attention for multi-speaker ASR in multi-party meeting scenario[C]//Proceedings of 2022 IEEE Spoken Language Technology Workshop (SLT). Doha: IEEE, 2023: 144-151.

[173] KINDT S, THIENPONDT J, MADHU N. Exploiting speaker embeddings for improved microphone clustering and speech separation in ad-hoc microphone arrays[C]//Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes: IEEE, 2023: 1-5.

[174] WANG H, GUO P, LI Y, et al. ICMC-ASR: The ICASSP 2024 in-car multi-channel automatic speech recognition challenge [EB/OL]. (2024-02-21). https://arxiv.org/html/2401.03473v3

[175] WU M, XU L, ZHANG J, et al. The USTC-nercslip systems for the ICMC-ASR challenge[C]//Proceedings of 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Seoul: IEEE, 2024: 3-4.

[176] HORIGUCHI S, TAKASHIMA Y, GARCÍA P, et al. Multi-channel end-to-end neural diarization with distributed microphones[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022: 7332-7336.

[177] PRASHANTH A, JAYALAKSHMI S L, VEDHAPRIYAVADHANA R. A review of deep learning techniques in audio event recognition (AER) applications[J]. Multimedia Tools and Applications, 2024, 83(3): 8129-8143.

[178] 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, 2021, 29: 684-698.

[179] MEYER M, CAVIGELLI L, THIELE L. Efficient convolutional neural network for audio event detection[EB/OL]. (2017-09-28). https://doi.org/10.48550/arXiv.1709.09888.

[180] 阎鹏. 2022年北京冬奥会雪上项目赛场内的拾音设计[J]. 演艺科技, 2022(1): 26-30, 50.
YAN Peng. The sound pickup design in the snow sports field of XXIV olympic winter games[J]. Entertainment Technology,

2022(1): 26-30, 50.

[181] 科大讯飞. 魔飞家居物联平台[EB/OL]. [2024-07-30]. https://morfeilink.himorfei.com.
iFlytek. MOFEI home-living IOT platform [EB/OL]. [2024-07-30]. https://morfeilink.himorfei.com.

[182] WANG Q, GUO S, YIU K F C. Distributed acoustic beamforming with blockchain protection[J]. IEEE Transactions on Industrial Informatics, 2020, 16(11): 7126-7135.

[183] ZENG Y, HENDRIKS R C. Distributed estimation of the inverse of the correlation matrix for privacy preserving beamforming [J]. Signal Processing, 2015, 107: 109-122.

[184] LI Q, HEUSDENS R, CHRISTENSEN M G. Privacy-preserving distributed optimization *via* subspace perturbation: A general framework[J]. IEEE Transactions on Signal Processing, 2020, 68: 5983-5996.

[185] HIMAWAN I, MCCOWAN I, SRIDHARAN S. Clustered blind beamforming from Ad-hoc microphone arrays[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(4): 661-676.

[186] GERGEN S, NAGATHIL A, MARTIN R. Classification of reverberant audio signals using clustered Ad-hoc distributed microphones[J]. Signal Processing, 2015, 107: 21-32.

[187] KINDT S, THIENPONDT J, BECKER L, et al. Robustness of Ad-hoc microphone clustering using speaker embeddings: Evaluation under realistic and challenging scenarios[J]. EURASIP Journal on Audio, Speech, and Music Processing, 2023, 2023(1): 46.

作者简介:

张结(1990-),**通信作者**,男,副研究员,硕士生导师,研究方向:语音信号处理、分布式麦克风阵列、类脑听觉感知,E-mail:jzhang6@ustc.edu.cn。

凌震华(1979-),男,教授,博士生导师,研究方向:语音信号处理、语音合成、自然语言处理,E-mail:zhling@ustc.edu.cn。

呼德(1993-),男,研究员,博士生导师,研究方向:分布式麦克风阵列、多媒体信息处理,E-mail:cshood@imu.edu.cn。

张晓雷(1983-),男,教授,博士生导师,研究方向:音频及语音信号处理、机器学习、人工智能,E-mail:xiaolei.zhang@nwpu.edu.cn。

(编辑:夏道家)