基于软标签编码的自组织麦克风阵列端到端声源定位

冯林峰、*Yijun Gong*、*Xiao-Lei Zhang*

海洋科学与技术学院，西北工业大学，中国，西北工业大学深圳研究院，中国
fenglinfeng@mail.nwpu.edu.cn，gongyj@mail.nwpu.edu.cn，xiaolei.zhang@nwpu.edu.cn

arXiv:2304.07512v1 [eess.AS] 15 Apr 2023

摘要

近年来，一种基于临时麦克风阵列的端到端二维声源定位算法将声源定位问题表述为分类问题。该算法将目标室内空间划分为一组局部区域，并预测说话人所在的局部区域。然而，局部区域采用独热编码表示，这可能会因量化误差而丢失局部区域之间的关联。本文提出一种新的软标签编码方法，称为标签平滑，用于基于临时麦克风阵列的分类式二维声源定位。其核心思想是在标签编码过程中纳入类别之间的几何关联。第一种方法称为静态软标签编码（SSLC），根据局部区域之间的距离将独热编码转换为软编码。由于 SSLC 是手工设计的，可能并非最优，因此第二种方法称为动态软标签编码（DSLC），它进一步修正 SSLC，在训练阶段根据分类式定位模型生成的预测结果的统计信息学习软编码。实验结果表明，所提出的方法能够有效提高定位精度。

索引词— 临时麦克风阵列、声源定位、软标签编码

1. 引言

声源定位是一个估计声源相对于麦克风阵列位置的问题。它有广泛的应用，例如语音增强与分离 [1], 语音识别 [2], 助听器 [3], 等。传统上，该问题被简化为估计声源的到达方向（DOA）。代表性方法包括到达时间差（TDOA） [4], 相位变换引导响应功率（SRP-PHAT） [5], 和多重信号分类（MUSIC）[6]。近年来，使用深度神经网络（DNN）进行声源定位的研究兴趣日益浓厚。这类方法将相位谱图 [7], 互相关 [8], 或空间伪谱 [9],等空间特征作为 DNN 的输入，以估计 DOA。上述方法均基于单个麦克风阵列。

如果联合使用多个麦克风阵列进行声源定位，就有可能直接获得声源的二维（2D）定位。自组织麦克风阵列通过在空间中协同组织一组随机分布的麦克风阵列，为解决这一问题提供了一种方案。早期方法大多基于传统信号处理方法 [10]。近年来，研究者开始探索基于深度学习的方法 [11], ，这也是本文的重点。尽管 [12] 中的开创性工作为这一领域作出了重要贡献，但这些工作对自组织麦克风阵列作出了严格假设，例如在训练和测试阶段麦克风节点的位置固定，因此未能充分发挥自组织阵列灵活性的优势。虽然 [13] 中的方法能够处理由任意数量、随机分布的临时节点构成的灵活自组织阵列，但它采用分阶段方法。其深度模型仅用于各个临时节点，使二维定位过程仍是一个独立的几何方法。

众所周知，深度学习的一项关键优势在于能够以端到端方式处理任务。在 [14], 中，提出了一种基于自组织阵列的端到端二维声源定位模型。该模型将二维声源定位问题构建为分类问题，其中每个类别代表目标房间中的一个局部区域，并以独热编码表示。此外，该方法适用于每个自组织节点仅包含一个麦克风的情况。然而，说话人的位置被估计为独热编码，这可能会带来问题。对于独热编码而言，正确类别与任一错误类别的距离都相同。

为解决这一问题，我们提出一种标签平滑算法，以弥补独热编码的不足。具体而言，我们首先提出基于

Xiao-Lei Zhang 为通讯作者。本研究部分由中国深圳市科技创新委员会项目资助，项目编号为 JSGG20210802152546026 和 JCYJ20210324143006016；另有部分由国家自然科学基金委员会（NSFC）资助，项目编号为 62176211。

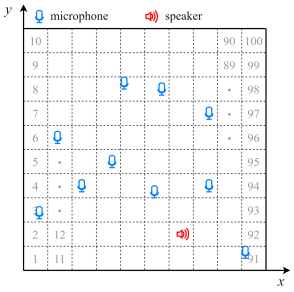


图 1。采用自组织麦克风阵列进行二维声源定位的分类式表述，其中房间被划分为多个局部区域。

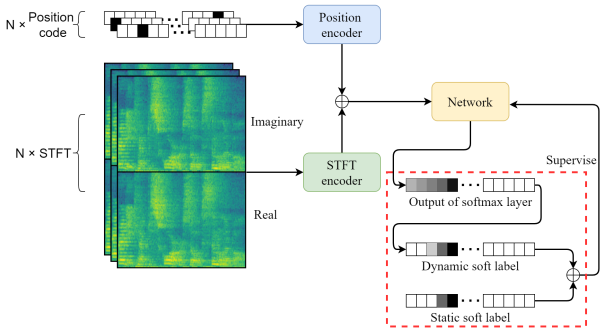


图2。所提出的采用软标签的端到端声源定位模型。红框中突出显示的是软标签生成模块。

局部区域中心之间距离的静态软标签编码（SSLC），以替代独热编码。由于 SSLC 是人工设计的，可能不够准确，因此我们进一步提出动态软标签编码（DSLC）来改进 SSLC。DSLC 根据定位模型在训练阶段产生的推理统计信息进行学习。实验结果证明了所提方法的有效性。

2. 相关工作

本文提出的方法与独热编码的局限性有关。具体而言，独热编码将任意一对类别之间的距离设为相同，且不考虑类别间的相似性。为解决这一问题，标签平滑 最初是为计算机视觉提出的 [15]，可提升分类模型的泛化性能。然而，它仍忽略了类别间的相似性。 [16]提出了一种基于预测结果统计信息的在线标签平滑方法。生成的软标签能够反映类别间的相似性。然而，将上述方法直接应用于声源定位会导致分类模型无法收敛，且性能较差。本文提出联合优化上述两种软标签，以解决端到端声源定位中独热编码的局限性。

3. 方法

3.1. 端到端声源定位

如图1所示，采用临时麦克风阵列进行端到端声源定位时，会将目标房间空间划分为若干局部区域，并用独热编码对每个区域进行编码。该方法的架构如图2所示（不包括红框中的模块），将临时节点位置的独热编码作为空间输入特征，并将来自临时节点的短时傅里叶变换（STFT）作为声学特征。该方法将声源定位问题表述为一个分类问题，旨在将声源归入某一个网格。由于空间信息编码在独热编码中，因此该方法适用于一种特殊场景：每个临时节点仅包含一个麦克风，这也是本文的工作场景。

原始的端到端声源定位采用独热编码对房间空间进行编码。假设一个包含 n个类别的分类问题所对应的独热编码集合为 $\mathcal{S}^{1-\text{hot}} = \{\mathbf{S}_i^{1-\text{hot}}\}_{i=1}^{n}$，其中第 i类的编码$\mathbf{S}_i^{1-\text{hot}} = [S_{i,1}^{1-\text{hot}}, \ldots, S_{i,k}^{1-\text{hot}}, \ldots, S_{i,n}^{1-\text{hot}}]$ 定义为：

$$S_{i,k}^{1-\text{hot}} = \begin{cases} 1, & \text{if } k = i \\ 0, & \text{otherwise} \end{cases}, \quad \forall k = 1, \ldots, n \tag{1}$$

其中 n 是局部区域的总数。

本节接下来将重点介绍图2红框中突出显示的新模块。

3.2. 静态软标签编码

与传统分类问题不同，传统分类问题中的类别可能本质上彼此不同且互不相关，而我们的声源定位问题中的类别彼此相关。例如，我们知道局部区域的几何位置及其相似性，例如任意两个局部区域中心之间的距离。因此，独热编码可能不适用于我们的问题。

在此，我们提出 SSLC 来解决这一问题。对于第 i类，记为 $\mathbf{S}_i^{\text{SSLC}} = [S_1^{\text{SSLC}}, \ldots, S_i^{\text{SSLC}}, \ldots, S_n^{\text{SSLC}}]$，SSLC 根据第 i个和第 k个局部区域中心之间的距离来设定其第 k个元素。形式上，SSLC 定义为距离的高斯函数：

$$S_{i,k}^{\text{SSLC}} = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{-d_{i,k}} e^{-\frac{\theta^2}{2\sigma^2}} d\theta, & \text{if } k \neq i \\ 1 - \sum_{\forall 1 \leq j \leq n, j \neq i}^{n} S_{j,k}, & \text{otherwise} \end{cases} \quad \forall k = 1, \ldots, n \tag{2}$$

其中，$d_{i,k}$ 是第 i个与第 k个局部区域之间的距离，σ 是控制高斯分布扩散范围的标准差。我们设定 $\sigma = l_{\text{ave}}/\alpha_s$，其中 l_{ave} 是

训练集中所有房间的局部区域对角线长度的平均值，α_s 是一个可调超参数。

根据上述公式，我们可以使用 SSLC，即 $\{\mathbf{S}_i^{\mathrm{SSLC}}\}_{i=1}^n$，作为真实标签而非独热编码来训练分类模型。单个话语的训练损失为：

$$L^{\mathrm{SSLC}} = -\sum_{k=1}^{n} S_{i,k}^{\mathrm{SSLC}} \log \hat{y}_k \tag{3}$$

其中，$\boldsymbol{x}$ 是模型的输入特征，$\hat{}y_k$ 是第 k 个单元的 Softmax 输出。

3.3. 动态软标签编码

SSLC 是一种人工设计标签，可能无法充分反映局部区域之间的相关性。为解决这一问题，我们在训练阶段通过在线标签平滑 [16] 来学习一个名为 DSLC 的编码簿。

DSLC 是在训练阶段借助真实独热标签得到的。具体而言，对于第 t 个训练轮次（$t \geq 1$），假设分类模型对输入话语的 Softmax 输出为 $\hat{}\mathbf{y}_j = [\hat{}y_1, \ldots, \hat{}y_n]$，该话语说话人位置的真实独热编码为 $\mathbf{S}_i^{1-\mathrm{hot}}$。如果正确预测了说话人位置，即 $\arg\max \hat{}\mathbf{y} = i$，则我们将 $\hat{}\mathbf{y}$ 选入一个记为 $\mathcal{Y}_i^{(t)}$ 的集合；否则，丢弃 $\hat{}\mathbf{y}$。

然后，我们通过对 $\mathcal{Y}_i^{(t)}$ 中的元素求平均，为第 i类生成一个 DSLC：

$$\mathbf{S}_i^{(t)} = \frac{1}{\left|\mathcal{Y}_i^{(t)}\right|} \sum_{\forall \hat{\mathbf{y}} \in \mathcal{Y}_i^{(t)}} \hat{\mathbf{y}}, \quad \forall i = 1, \ldots, n \tag{4}$$

最后，对于第（$t+1$）个训练轮次，我们使用在第 t个训练轮次生成的 DSLC，即 $\{\mathbf{S}_i^{(t)}\}_{i=1}^n$，作为监督信号来指导训练，其中，本训练轮次中单个话语的训练损失定义为：

$$L_{(t+1)}^{\mathrm{DSLC}} = -\sum_{k=1}^{n} S_{i,k}^{(t)} \log \hat{y}_k \tag{5}$$

其中，$S_{i\,k}^{(t)}$,是$\mathbf{S}_i^{(t)}$ 的第 k个元素。

需要注意的是，在第一个训练轮次之前，即 $t = 0$，我们使用标准标签平滑 [15] 来初始化 $\{\mathbf{S}_i^{(0)}\}_{i=1}^n$。

3.4. 联合训练

第（$t+1$）个训练轮次中 DSLC 的可靠性取决于第 t个训练轮次中的深度模型。由于初始深度模型过于薄弱，从一开始就训练 DSLC 非常困难，有时甚至无法收敛。

为解决这一问题，我们提出了一种联合训练策略，使用静态标签（如 SSLC 或独热编码）来辅助 DSLC 的训练：

$$L_{(t)}^{\mathrm{joint}} = \alpha_d^{(t)} L_{(t)}^{\mathrm{DSLC}} + (1 - \alpha_d^{(t)}) L^{\mathrm{SSLC}} \tag{6}$$

其中，$\alpha_d^{(t)}$ 是第 t训练轮次的可调超参数。

这引出了如何设置 $\alpha_d^{(t)}$ 的问题。第一种选择是将 $\alpha_d^{(1)} = \alpha_d^{(2)} = \ldots = \alpha_d^{(T)}$ 设为常数 α_d。另一种选择是根据 $L_{(t)}^{\mathrm{DSLC}}$的重要性自适应地调整 $\alpha_d^{(t)}$。本文中，我们将 $\alpha_d^{(t)} =$ 设为 $\mathrm{ACC}^{(t)}$，其中 $\mathrm{ACC}^{(t)}$ 是第 t个模型在训练集上的预测准确率。当然，我们也可以将 $\alpha_d^{(t)}$ 设为深度模型的可学习参数。

4. 实验

4.1. 实验设置

我们在模拟数据上进行了实验。源语音来自 LibriSpeech 语料库 [17]。该语料库的 train-clean-360 子集、dev-clean 子集和 test-clean 子集分别包含 921、40 和 40 位说话人。我们使用 Torchaudio 模块 [18] 提供的语音活动检测（VAD）来移除静音片段。VAD 后时长不足 2 秒的语音片段会被丢弃。对于其余每个片段，我们随机选取一段 2 秒的音频，以生成带混响的多通道数据。我们使用 Pyroomacoustics [19] 模块生成房间脉冲响应。所有模拟语音的混响时间 T_{60} 均服从从$[0.3, 1.0]$ 秒起的均匀分布。我们还在混响语音中加入了加性噪声。加性噪声从一个大型噪声集中随机选取，该噪声集包含 126 小时的不同类型噪声片段。每段话语的信噪比从 $[20, 50]$ dB 的范围内随机选取。最终，训练集、验证集和测试集中的话语数量分别为 18000、2500 和 2500。

我们生成了 15 个不同大小的房间，其中 10 个用于训练，5 个用于测试。房间的长和宽范围为 4 米至 10 米。房间高度固定为 3 米。麦克风阵列和声源的高度固定为 1 米。每个房间在二维平面上被划分为 15×15 个局部区域。对于每段话语，我们首先随机选择一个房间，然后在 225 个局部区域之一随机设置语音声源，最后在其余 224 个局部区域中随机放置 30 个自组织麦克风节点。

我们使用 [14]中的深度神经网络模型作为骨干网络。首先，它分别将临时布设节点的位置编码和短时傅里叶变换声学特征作为两个编码器的输入。随后，将两个编码器的输出拼接后，送入时空注意力网络。该注意力网络依次执行跨通道注意力、通道融合和跨帧注意力。说话人的局部区域被视为时空注意力网络沿时间轴的平均输出。我们使用 ReLU 作为激活函数，并使用 Adam 作为优化器来训练模型。

表1。所提出的五种软标签编码方法和独热编码的平均绝对误差（米）及准确率（%）结果。

方法	房间1		房间2		房间3		房间4		房间5		平均值	
	平均绝对误差	准确率	平均绝对误差	准确率	平均绝对误差	准确率	平均绝对误差	准确率	平均绝对误差	准确率	平均绝对误差	准确率
独热编码 [14]	0.2135	54.6	0.2471	51.0	0.2959	46.6	0.3283	**49.8**	0.3207	54.0	0.2811	51.2
SSLC	0.2260	53.4	0.2566	49.2	0.3001	**48.8**	0.3363	43.6	0.3344	52.8	0.2907	49.6
DSLC+独热编码（α_d 为常数）	0.2226	54.2	0.2592	47.6	0.3079	43.0	0.3234	47.4	0.3225	55.4	0.2871	49.5
DSLC+SSLC（α_d 为常数）	0.2115	**57.6**	**0.2448**	**52.4**	**0.2873**	47.0	0.3088	48.2	**0.3185**	**56.0**	**0.2742**	**52.2**
DSLC+独热编码（$\alpha_d^{(t)}$ 为变量）	0.2149	56.0	0.2689	45.0	0.3052	45.0	0.3239	49.6	0.3408	47.8	0.2907	48.7
DSLC+SSLC（$\alpha_d^{(t)}$ 为变量）	**0.2059**	57.2	0.2577	47.8	0.2974	47.8	**0.2998**	47.4	0.3351	53.2	0.2792	50.7

表2。独热编码与所提出的 DSLC+SSLC（α_d 为常数）方法的学习误差（单位：米）比较。

Room	独热编码	DSLC+SSLC	相对减少量（%）
房间1	0.0701	0.0681	2.83
房间2	0.0942	0.0919	2.44
房间3	0.1310	0.1224	6.56
房间4	0.1484	0.1289	13.14
房间5	0.1156	0.1134	1.90
平均值	0.1119	0.1049	6.26

我们使用平均绝对误差（**平均绝对误差**）评估所提出的方法:

$$\mathrm{MAE} = \frac{1}{I}\sum_{i=1}^{I}\sqrt{\left(x_i^{spkr} - \tilde{x}_i^{spkr}\right)^2 + \left(y_i^{spkr} - \tilde{y}_i^{spkr}\right)^2} \tag{7}$$

其中， I 是测试语句的数量，(x_i^{spkr} ， y_i^{s})是第 i个说话人位置的真实坐标，而($\tilde{x}_i^{spkr}$ ， $\tilde{y}_i^{spkr}$)是预测说话人位置 $\tilde{y}_i$所在局部区域中心的坐标。

由于局部区域内的区域被量化为局部区域中心，因此基于声源定位分类形式的任何方法都不可避免地存在**量化误差**，它构成了误差上界。我们将该上界记为 UB-MAE。本文中，5个测试房间的 UB-MAE 平均值为 0.1693m。

4.2. 结果

我们采用 6 种标签编码方法训练了声源定位模型，如表1所示。从结果可以看出，所提出的联合训练方法 DSLC+SSLC（α_d为常量）总体上优于独热编码。在大多数情况下，准确率越高，平均绝对误差越小。无论 $\alpha_d^{(t)}$ 是常量还是变量，联合训练的性能始终优于独热编码。自适应调整 α_d 的方案并未比将 α_d 设为常量的方案取得更好的结果。这可能是因为将 $\alpha_d^{(t)} =$ 准确率$^{(t)}$ 设定为常量会使 DSLC 在训练集上过拟合。

表3。SSLC 中 α_s 对性能的影响。

α_s	2.6	2.7	2.8	2.9	3
平均绝对误差 (m)	0.2793	0.2796	0.2778	0.2782	0.2822
准确率 (%)	50.5	51.7	51.0	51.6	50.7

表4。“DSLC+SSLC（α_d 为常数）”中 α_d 对性能的影响。

α_d	0.2	0.4	0.5	0.6	0.8
平均绝对误差（m）	0.2839	0.2741	0.2726	0.2772	0.2845
准确率（%）	49.2	50.6	51.4	49.6	50.2

由于量化误差始终存在，表1中的误差包含两部分：一部分来自量化误差，另一部分来自编码方法导致的学习误差。我们在表2中进一步单独考察了学习误差，发现所提出的方法使其平均值相对降低了 6.26%。

SSLC 中有两个可调超参数 α_s ，以及“DSLC+SSLC（α_d 为常数）”中的α_d 。这里我们研究它们对性能的影响。对于 SSLC 中的 α_s ，当 α_s <2。6 时，说话人位于真实局部区域内的概率低于 0.8，导致性能较差；当 $\alpha_s > 3.0$ 时，SSLC 非常接近独热编码。因此，我们在表3中研究了 $\alpha_s \in [2.6, 3.0]$ 。从表中可以看出，α_s 的最优值为 2.8。对于“DSLC+SSLC（α_d 为常数）”中的 α_d ，我们在（0，1）范围内搜索了 α_d，结果发现，如表4所示， α_d 的最优值为 0.5。

5. 结论

本文针对自组织麦克风阵列端到端声源定位中独热编码的不足，提出了一种标签平滑算法。该算法包括一种根据局部区域中心距离计算的人工 SSLC，以及一种可学习的 DSLC，后者在训练过程中利用正确分类样本统计进行更新。实验结果表明，与独热编码相比，所提出的算法效果更佳，且不会增加模型复杂度和训练时间。

6. 参考文献

[1] 王德亮和陈吉桐，“基于深度学习的监督语音分离：综述”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 26 卷，第 10 期，第 1702–1726 页，2018 年。

[2]康斯坦丁·施皮勒、比尔格·科尔迈尔和贝恩德·T·迈耶，《简单与复杂声学场景中的人类语音识别与自动语音识别比较》，*ComputerSpeech &Language*，第52卷，第123–140页，2018年。

[3]蒂姆·范登博加特、埃夫琳·卡雷特和扬·沃特斯，《使用麦克风分别置于耳后式、耳道内式和耳廓内式位置的助听器进行声源定位》，《国际听力学杂志》，第50卷，第3期，第164–176页，2011年。[4] 查尔斯·纳普和格利福德·卡特，《用于时延估计的广义相关法》，*IEEE*《声学、语音与信号处理汇刊》，第24卷，第4期，第320–327页，1976年。[5] 约瑟夫·H·迪比亚斯、哈维·F·西尔弗曼和迈克尔·S·布兰德斯坦，《混响房间中的鲁棒定位》，载于麦克风阵列，第157–180页。施普林格，2001年。[6] 拉尔夫·施密特，《多发射源定位与信号参数估计》，*IEEE*《天线与传播汇刊》，第34卷，第3期，第276–280页，1986年。[7] 苏米特罗·查克拉巴蒂和埃马努¨埃尔·AP·哈贝茨，《采用噪声信号训练的深度卷积网络进行多说话人到达方向估计》，*IEEE*《信号处理选题汇刊》，第13卷，第1期，第8–21页，2019年。[8] 肖雄、赵盛奎、钟雄虎、道格拉斯·L·琼斯、庄荣祥和李海洲，《一种基于学习的方法：在噪声和混响环境中进行到达方向估计》，载于*2015*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2015年，第2814–2818页。[9] 阮氏玉寿、颜文成、里沙布·兰詹和道格拉斯·L·琼斯，《利用空间伪谱和卷积神经网络进行鲁棒声源计数与到达方向估计》，*IEEE/ACM*《音频、语音与语言处理汇刊》，第28卷，第2626–2637页，2020年。

[10]安东尼·格里芬、阿纳斯塔西奥斯·亚历山德里迪斯、De- spoina Pavlidi、扬尼斯·马斯托拉基斯和阿萨纳西奥斯·穆赫塔里斯，《在无线声学传感器网络中定位多个音频声源》，《信号处理》，第107卷，第54–67页，2015年。

[11]Manuel Hahmann、Efren Fernandez-Grande、Henrry Gunawan和Peter Gerstoft，《使用多个分布式自组织麦克风阵列进行声源定位》，*JASAExpressLetters*，第2卷，第7期，页码074801，2022年。[12] Fabio Vesperini、Paolo Vecchiotti、Emanuele Principi、Stefano Squartini和Francesco Piazza，《使用深度神经网络在多个房间中定位说话人》，*Computer Speech & Language*，第49卷，第83–106页，2018年。[13] Shupei Liu、Yijun Gong和Xiao-Lei Zhang，《基于深度学习的大型临时麦克风阵列二维说话人定位》，*arXiv*预印本 *arXiv:2210.10265*，2022年。[14]Yijun Gong、Shupei Liu和Xiao-Lei Zhang，《使用临时麦克风阵列进行端到端二维声源定位》，*arXiv*预印本 *arXiv:2210.08484*，2022年。[15] Christian Szegedy、Vincent Vanhoucke、Sergey Ioffe、Jon Shlens和Zbigniew Wojna，《重新思考用于计算机视觉的Inception架构》，载于*IEEE*计算机视觉与模式识别会议论文集，2016年，第2818–2826页。[16] Chang-Bin Zhang、Peng-Tao Jiang、Qibin Hou、Yunchao Wei、Qi Han、Zhen Li和Ming-Ming Cheng，《深入探究标签平滑》，*IEEE Transactions onImage Processing*，第30卷，第5984–5996页，2021年。[17] Vassil Panayotov、Guoguo Chen、Daniel Povey和Sanjeev Khudanpur，《LibriSpeech：基于公有领域有声读物的自动语音识别语料库》，载于*2015*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2015年，第5206–5210页。[18]Yao-Yuan Yang、Moto Hira、Zhaoheng Ni、Artyom Astafurov、Caroline Chen、Christian Puhrsch、David Pollack、Dmitriy Genzel、Donny Greenberg、Edward Z Yang等，《TorchAudio：音频与语音处理的构建模块》，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第6982–6986页。[19] Robin Scheibler、Eric Bezzam和Ivan Dokmani´c，《Pyroomacoustics：用于音频房间仿真和阵列处理算法的Python软件包》，载于*2018*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018年，第351–355页。