arXiv:2302.01483v1 [cs.LG] 3 Feb 2023

SPADE：自监督用于声学解耦的预训练

John Harvill[1*], *Jarred Barber*[2†], *Arun Nair*[2], *Ramin Pishehvar*[2]

[1]伊利诺伊大学厄巴纳-香槟分校，美国[2]亚马逊
Alexa AI，美国

摘要

自监督表征学习方法日益受到欢迎，因为它们能够利用大量无标签数据训练模型，并已在自然语言处理、计算机视觉和语音等多个领域取得成功。此前，语音领域的自监督研究已对语音的多种属性进行了解耦，例如语言内容、说话人身份和韵律。在本研究中，我们提出一种自监督方法，将房间声学与语音解耦，并将声学表征用于下游的设备仲裁任务。结果表明，当有标签训练数据稀缺时，我们提出的方法相比基线显著提升了性能，这说明我们的预训练方案能够学习编码房间声学信息，同时对语音信号的其他属性保持不变。

索引词— 关键词检出、声源定位、自监督预训练、解耦、声学

1. 引言

将语音解耦为多个组成部分，是信号处理中的一个基本问题，其应用包括语音转换 [1, 2, 3, 4], 自动语音识别 [5, 6, 7, 8], 说话人识别 [9], 以及语音隐私保护 [10]。解耦的目标是分离语音信号的不同属性，使最终的信号表征不受与下游任务无关的属性影响。以往的研究主要关注对声学内容的不变性[1]，从而突出说话人身份或语言内容等目标属性。本文则特别希望保留声学内容，并以自监督的方式提取对所有其他属性均具有不变性的表征。随后，这些表征将用于设备仲裁任务 [11]。

由于支持智能语音助理的设备无处不在，设备仲裁任务近年来应运而生。我们将这类设备统称为“语音助理”（虚拟助理，VA）或“设备”。如今，许多家庭在同一个房间里拥有多个虚拟助理，因此难以判断应由哪台设备与用户交互。用户想要开始与虚拟助理交互时，必须先说出唤醒词（“Alexa”、“Hey Google”等），房间内的所有设备都会录下这段语音。为了提供最自然的用户体验，只有用户意图使用的设备应被唤醒并继续与用户交互。这便引出了设备仲裁问题：

给定 N 段源音频录音，其中每段录音均来自一台虚拟助理，确定用户意图使用的是哪台虚拟助理。

请注意，设备仲裁与经过充分研究的声源定位问题密切相关 [12]。到达时间差（TDOA） [13] 是一种有效的技术，可用于解决音频信号的声源定位问题，但遗憾的是，它需要使用虚拟助理上没有配备的大型麦克风阵列。现代虚拟助理还受到若干其他限制，使设备仲裁成为一个并非易事的问题：(1) 每台虚拟助理的位置未知，而且可能随时间变化（例如将设备从客厅移到厨房）。(2) 虚拟助理所处的声学环境未知。(3) 设备之间不一定能够实现时钟同步，因此无法在设备之间使用 TDOA。鉴于这些限制，设备仲裁任务必须仅依靠每段音频录音中包含的房间声学信息来解决。

受这些限制的启发，我们提出了 SPADE：用于声学解耦的自监督预训练。SPADE 是一种预训练技术，无需标签，即可利用源音频的多个视图，从语音信号中解耦声学信息。我们发现，将 SPADE 与以往的设备仲裁研究结合使用时 [11], , 在有标签训练数据较少的情况下，能够提升性能。鉴于此任务的有标签数据难以收集，SPADE 是一种极具价值的技术，无需增加任何推理成本，即可提升设备仲裁的性能。

本文其余部分安排如下：第 2 节讨论设备仲裁及相关任务的先前研究。第 3 节介绍实验所用的数据，以及生成这些数据的仿真过程。第 4 节详细介绍我们提出的预训练方法。第 5 节讨论实验设置，第 6 节讨论结果。最后，第 7 节总结全文并提出未来研究方向。

2. 先前研究

尽管设备仲裁是一个相对较新的问题，但它与声源定位 [12, 13] 密切相关。声源定位的目标是确定发出声音的物体（即声源）的位置。当前技术依赖于大型麦克风阵列，而在我们的问题设置中无法使用这类阵列。先前研究 [11]还表明，直接预测各设备与声源之间的距离，并据此做出仲裁决策，会导致性能下降。

仲裁的目标是根据距离和方向，选择用户正在与之交谈的被关注设备。常见技术会比较信号属性，例如信噪比（SNR）、声源与麦克风之间的估计距离、互相关等 [14]。设备仲裁的目标不同于通道选择 [14, 15] ，因为最优设备的定义仅仅是距离用户最近的设备，或被关注设备（定义为

*本研究是在亚马逊 Alexa AI 实习期间完成的。† 现就职于 Google Research[1]本文中的“声学内容”或“声学”指音频信号中与特定位置的房间脉冲响应（RIR）相关的信息；在该位置播放源音频时，音源位于房间内另一个固定位置。

参数	分布
房间长/宽（米）	均匀分布(3.0, 10.0)
房间高度（米）	均匀分布(2.5, 6.0)
混响时间 (s)	Beta(1.1, 3.0)
设备数量	平移泊松分布(m=3,l=2,h=15)
噪声源数量	平移泊松分布(m=2,l=1,h=5)
语音级（声压级分贝）	均匀分布(55.0, 70.0)
噪声级（声压级分贝）	均匀分布(50.0, 70.0)

表1。场景采样的超参数。对于 ShiftedPoisson 分布，我们用“m”表示均值，用“l”表示下限，用“h”表示上限。

用户视线方向上的设备，或根据其他相关的声学或视觉线索确定的设备），而不是能够在其他下游任务中带来更好性能的设备。设备仲裁与声源定位在问题设置和目标上的差异表明，设备仲裁应当作为独立课题进行研究，并且具有自身独特的一系列挑战，因此是一个值得探讨的问题。

我们采用了 Barber 等人提出的基于机器学习的设备仲裁方法作为基线 [11]。作者提出了一种端到端方法，用于训练神经网络执行设备仲裁。其模型由一个在每台设备上本地运行的小型卷积特征提取器，以及一个在云端运行、用于做出最终决策的较大型仲裁网络组成。结果表明，与简单的基于能量的方法相比，该方法在各种房间条件下的性能均显著提升。

3. 数据集模拟

目前，尚无包含已知真实标签的大规模设备仲裁数据集，因此我们按照 Barber 等人 [11] 描述的三个主要步骤，使用模拟数据进行实验：

- **采样场景：** 从各种声学环境设置（房间长/宽、噪声源）中采样一个房间，并采样房间内的设备/扬声器位置。给定设备/扬声器位置，根据设备与扬声器之间的最小欧几里得距离生成仲裁标签。
- **生成 RIR：** 根据采样得到的场景，使用声学模拟器为场景中的每台设备生成房间脉冲响应（RIR）。
- **生成音频：** 将语音片段与生成的 RIR 进行卷积，并为每台设备混入噪声。完成此步骤后，我们便得到一个包含设备仲裁音频及其对应真实标签的人工数据集。

我们使用镜像源法（ISM） [16] 进行数据模拟，并使用来自 [11]，的源音频，但更新了场景采样
与表1中的超参数相同。

4. 方法

我们的仲裁模型基于 Barber 等人提出的方法 [11]，由两个组件构成：特征编码器，以及用于作出仲裁决策的分类器。在对编码器和分类器进行端到端训练之前，我们采用两种方案对编码器进行预训练：对比式预训练和重构式预训练。下文各小节将介绍我们的预处理流程和预训练方案。

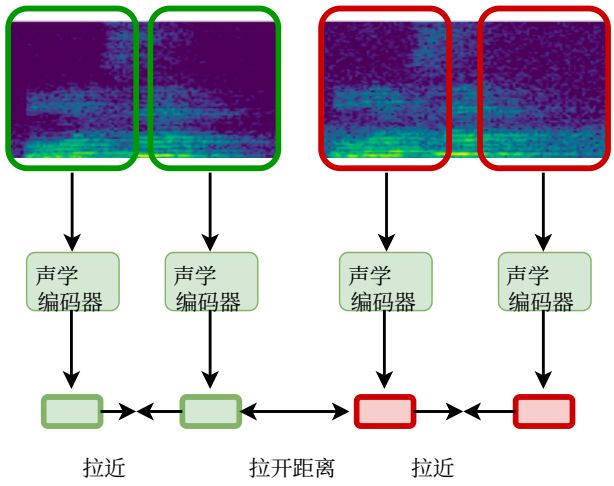


图 1. 对比预训练方案：声学编码器共享权重。

4.1. 音频预处理

首先将音频转换为对数滤波器组能量（LFBE）特征，频谱图采用 25ms 的帧长和 10ms 的帧移计算。对频谱图应用包含 64 个梅尔频带的梅尔变换，随后进行对数变换。进入神经网络模型进行后续处理前，对 LFBE 特征进行均值和方差归一化。

4.2. 编码器架构

编码器架构采用残差卷积模型，由 18 个卷积层组成，每一层均应用批归一化和 ReLU 激活函数。本文讨论的所有方法（预训练、基线）使用相同的编码器模型。该网络生成输入的序列特征表示，其时间分辨率远低于输入的 LFBE 特征。预训练期间，我们使用一个小型 Transformer [17]网络，将编码器输出的向量序列映射为单个向量。预训练后会移除此网络，但在下文有关预训练的讨论以及图 1 和图 2 中，它隐含地作为编码器的一部分。

4.3. 对比式预训练

研究表明，对比损失函数能够在语音和自然语言处理等多个领域生成高质量的表征 [18, 19]。这类损失函数的工作原理是，假设数据在某个特定属性上相似或不同，并通过嵌入空间中的距离，促使数据嵌入反映这些关系。在设备仲裁场景中，来自同一设备的音频具有相同的房间声学属性，而来自不同设备的音频则具有不同的声学属性。[2] 由于我们希望编码房间声学信息，因此可以促使来自同一设备的音频嵌入彼此相似，同时促使来自不同设备的音频嵌入彼此正交（见图 1）。由于来自给定 N 的所有录音中的语音内容都相同，

[2]这是 由于假设这些设备位于不同的位置 。

因此，嵌入应不受语音内容影响，而仅编码声学信息。

将设备 a 录制的音频记为 x^a。对于这种预训练方法，我们将 x^a 输入编码器[3]，，即 z^a = encoder(LFBE(x^a))，从而得到定长嵌入 z^a 。然后将 z^a 归一化，使其幅值为 1。我们将 x^a 中从时间 t_1 到 t_2 的一段连续切片记为$x^a_{t_1:\ t_2}$，其中 $t_2 > t_1$；其对应的定长嵌入记为$z^a_{t_1:\ t_2}$。给定 N 段录音 x^1，x^2，..... x^N，将它们零填充至相同长度 T，并给定一个分割索引 t_{split}，满足 $0 < \frac{T}{2} - \epsilon < t_{split} < \frac{T}{2} + \epsilon < T$，其中 ϵ 是一个较小的随机抖动，我们得到单个仲裁场景的损失函数 $\mathcal{L}_C$：

$$\mathcal{L}_1 = \sum_{i=1}^{N}\sum_{j=1}^{N} |\langle z^i_{0:t_{split}}, z^j_{t_{split}:T}\rangle - \delta_{ij}| \tag{1}$$

$$\mathcal{L}_2 = \sum_{i=1}^{N}\sum_{j\neq i} |\langle z^i_{0:t_{split}}, z^j_{0:t_{split}}\rangle| + |\langle z^i_{t_{split}:T}, z^j_{t_{split}:T}\rangle| \tag{2}$$

$$\mathcal{L}_C = \mathcal{L}_1 + \mathcal{L}_2 \tag{3}$$

其中 δ_{ij} 是克罗内克δ函数。请注意，$\mathcal{L}_1$ 鼓励同一段音频录音的两个片段映射到相同的嵌入。$\mathcal{L}_2$ 项通过鼓励不同录音中各自对应的音频片段彼此正交，为语音内容不变性提供了更强的监督。

4.4. 重构式预训练

此前已有研究采用带信息瓶颈的自编码方法，实现了语音中不同属性的解耦 [1, 2]。我们采用类似的方法，以自监督的方式将房间声学信息与语音解耦，并假设在唤醒词音频的持续时间内（~2s），房间声学属性保持不变[4]。

我们的模型由语音编码器、声学编码器和重构解码器组成。语音编码器 $S(\cdot)$ 和重构解码器 $R(\cdot)$ 均为 Transformer [17]，且各自生成一个向量序列。鉴于声学的平稳性假设，以及创建信息瓶颈的需要，我们将声学编码器 $A(\cdot)$ 设计为生成一个固定维度的嵌入。由于同一源音频的 N 个录音包含相同的语音内容，因此每个录音的语音表征 $s_i = S(x^i)$ 应当相同，即 $s^1 \approx s^2 \approx ... \approx s^N$。为此，我们使用一个录音的声学嵌入和同一房间内另一个录音的语音嵌入，重构其中一个录音的 LFBE 特征。通过创建一个小型固定维度嵌入来施加信息瓶颈，可促使声学编码器仅表征所有信号中不共有的信息，即房间声学信息。由于这 N 个录音在时间上未对齐，我们通过提取 LFBE 特征的包络 $env(x^i)$ 向解码器提供对齐信息；该包络是每个时间步特征向量的均值：

$$env(x^i)_k = \frac{1}{N}\sum_{q=1}^{N} \text{LFBE}(x^i)_{kq} \tag{4}$$

[3]此处默认第 4.2 节讨论的 Transformer 网络位于编码器之后，用于生成固定长度表示。[4]这一平稳性假设并非在所有情况下都成立（例如人或宠物四处走动），但考虑到音频仅录制两秒，这一假设是合理的。

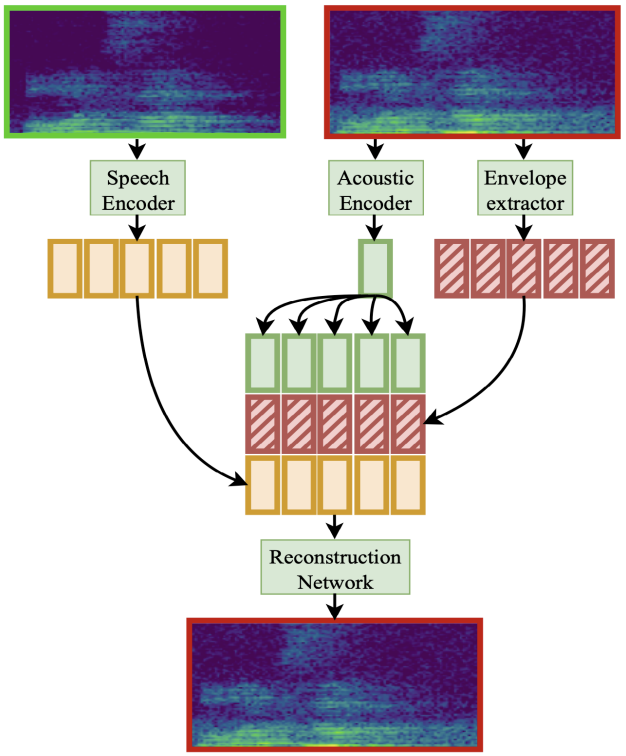


图 2。重建预训练方案：语音编码器、包络和重建网络的时间轴维度与 LFBE 时间轴相同。声学编码器生成固定长度嵌入，并沿语音编码器和包络输出的时间轴复制该嵌入。斜线阴影单元来自不可学习模块。

其中，k 表示时间轴，q 表示 LFBE 特征矩阵的特征轴，而 $N = 64$。请注意，包络提取器并非可学习模块，而是在每个时间步执行简单的均值运算，以生成 LFBE 特征的低维表示（特征向量 $\rightarrow$ 标量）。对于一个仲裁场景，我们可以将损失函数 $\mathcal{L}_R$ 形式化地写为：

$$\mathcal{L}_R = \frac{1}{N}\sum_{i=1}^{N} ||\text{LFBE}(x^i) - R(S(x^{\neq i}), A(x^i), env(x^i))||_2^2 \tag{5}$$

其中，$x^{\neq i}$ 表示从$\{x^1, x^2, ..., x^N\}$中随机选取的一段音频录音，且该录音不是 x^i。我们沿时间轴复制 $A(x^i)$，并将其与 $S(x^{\neq i})$和 $env(x^i)$ 拼接起来，形成 $R(\cdot)$的输入。图 2 展示了这一过程的示意图。

4.5. 仲裁分类器架构

此前，仲裁分类器采用多层感知机（MLP）网络实现 [11]，但我们发现，使用类似 Transformer [17] 的自注意力网络可以进一步提升性能。对于每个设备 i，编码器都会输出一系列隐藏状态 h^i_1，h^i_2，..... h^i_K。对于 N 个设备，将隐藏状态沿时间轴拼接，形成序列：

$$H = h^1_1, h^1_2, ..., h^1_K, h^2_1, h^2_2, ...h^2_K, ..., h^N_1, h^N_2, ...h^N_K \tag{6}$$

然后将该序列输入由自注意力层[5] 组成的网络，以生成以下序列：

$$G = g_1^1, g_2^1, ..., g_K^1, g_1^2, g_2^2, ...g_K^2, ..., g_1^N, g_2^N, ...g_K^N \quad (7)$$

随后，每个序列 $g_1^i, g_2^i, ..., g_K^i$ 都会经过第二个自注意力网络，以生成随时间变化的摘要 G_i。然后，每个 G_i都会经过一个两层前馈神经网络，输出设备 i对应的标量逻辑值。接着将这些逻辑值输入 Softmax层，得到仲裁概率。整个分类网络通过最小化仲裁概率与真实标签分布之间的交叉熵损失进行优化。

4.6. 基线

本文的贡献在于自监督预训练方法，因此我们将此前讨论的编码器和分类器网络用作基线，但不对编码器进行预训练。

5. 实验

训练数据集包含 30 万个仲裁场景。为了展示预训练在学习有助于设备仲裁的表征方面的有效性，我们构建了规模呈指数递减的数据集。对于数据集 i，训练集大小为 $s_i = \lfloor S/4^i \rfloor$，其中 S =300k（完整训练集的规模）。我们选择 $i \in \{0, 1, 2, 3\}$，使得最小训练集包含 $\sim$4.7k 个场景。

5.1. 实验流程

对于每项实验，我们根据验证损失最低的检查点选择最终的仲裁模型，并在保留测试集上进行评估。我们设置了四种实验方案：

- **基线：** 针对每个大小为 s_i 的训练数据子集，端到端训练编码器-分类器模型，训练 $i \in \{0, 1, 2, 3\}$。
- **对比式：** 使用对比式方法在所有可用训练数据（30 万个场景，不涉及标签）上预训练（声学）编码器。选择验证表现最佳的检查点来初始化编码器，然后针对每个大小为 s_i $i \in \{0, 1, 2, 3\}$ 的训练数据子集，端到端微调编码器-分类器模型。
- **重构式：** 与对比式方案相同，但我们使用重构式预训练。
- **组合：** 同时使用对比式预训练和重构式预训练。损失函数变为 $\mathcal{L} = \lambda\mathcal{L}_R + (1 - \lambda)\mathcal{L}_C$，其中我们设定 $\lambda = 0.5$。按照前述方案微调编码器-分类器模型。

6. 结果

此处以相对错误率 表示结果，该指标是相对于 4.7k 基线设置的性能而言的。将目标方法 m 的准确率记为 acc_m，将 4.7k 基线设置的准确率记为 acc_{base}，相对错误率的计算公式如下：

$$\text{err}_{\text{rel}} = \frac{1 - \text{acc}_m}{1 - \text{acc}_{\text{base}}} \quad (8)$$

[5]在输入中加入位置编码。

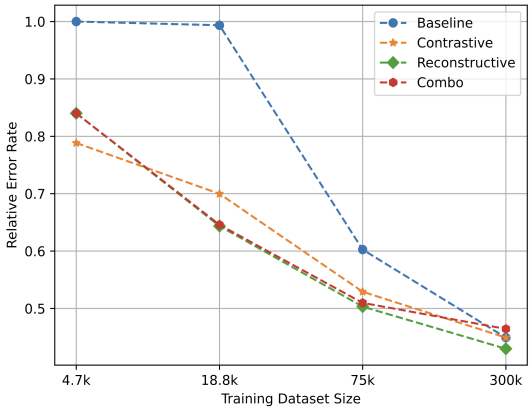


图 3。相对于最差情况的相对错误率。

我们的结果如图 3 所示。最重要的趋势是，训练数据集越小，预训练方法相较于基线的优势就越明显。这表明，我们提出的预训练方案能够保留声学信息，学习与设备仲裁问题相关的特征，并且在有标签训练数据稀缺时最为有益。使用预训练检查点初始化编码器，有助于缓解设备仲裁训练过程中的过拟合。我们还发现，对比式预训练与重构式预训练相结合，并未比单独使用任一种方法带来明显提升，这表明两种方法编码的信息可能相似。

7. 结论

本文提出了对比式和重构式预训练这两种自监督表征学习方法，用于从语音中分离声学内容。与以往旨在创建对声学内容保持不变的表征的工作不同，我们旨在仅编码声学内容，并通过设备仲裁问题展示其用途。我们发现，所提出的两种预训练方法都优于基线，而且当有标签训练数据集较小时，改进更为显著。这为我们的预训练目标提供了实证依据：即使缺少大型训练语料库，所得到的声学内容表征也可用于设备仲裁任务。

由于自监督技术不需要人工标注，因此我们提出的方法或许也能应用于其他研究问题。例如，我们的解耦声学表征或可用于其他声学任务，例如房间声学特性估计或家庭影院的声学自适应。虽然本文未对此进行研究，但通过重构式预训练学习到的语音内容表征，或许对说话人识别或自动语音识别等声学不变应用具有价值，因为这些表征旨在编码除声学信息以外的所有信息。

8. 参考文献

[1] Kaizhi Qian、Yang Zhang、Shiyu Chang、Xuesong Yang 和 Mark Hasegawa-Johnson，“Autovc：仅使用自编码器损失的零样本语音风格迁移”，载于国际机器学习会议。PMLR，2019年，第5210–5219页。[2] Kaizhi Qian、Yang Zhang、Shiyu Chang、Mark Hasegawa-Johnson 和 David Cox，“通过三重信息瓶颈进行无监督语音分解”，载于国际机器学习会议。PMLR，2020年，第7836–7846页。[3] Jie Wang、Jingbei Li、Xintao Zhao、Zhiyong Wu、Shiyin Kang 和 Helen Meng，“通过对抗式学习解耦语音表征，实现稳健的多因素语音转换”，*arXiv*预印本 *arXiv:2102.00184*，2021年。[4] Jiachen Lian、Chunlei Zhang 和 Dong Yu，“面向零样本语音转换的稳健解耦变分语音表示学习”，载于*ICASSP 2022—2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第6572–6576页。[5] Kaizhi Qian、Yang Zhang、Heting Gao、Junrui Ni、Cheng-I Lai、David Cox、Mark Hasegawa-Johnson 和 Shiyu Chang，“Contentvec：通过解耦说话人信息改进自监督语音表征”，载于国际机器学习会议。PMLR，2022年，第18003–18017页。[6] Janek Ebbers、Michael Kuhlmann、Tobias Cord-Landwehr 和 Reinhold Haeb-Umbach，“结合对比预测编码的因子化变分自编码器，用于无监督解耦语音表征学习”，载于*ICASSP 2021—2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2021年，第3860–3864页。[7] Wei-Ning Hsu、Hao Tang 和 James Glass，“利用可解释的解耦表示进行自适应，实现远场会话语音识别”，*I nterspeech*，2018年。[8] Sameer Khurana、Shafiq Rayhan Joty、Ahmed Ali 和 James Glass，“用于从语音中无监督学习解耦表示的因子化深度马尔可夫模型”，载于*IC ASSP 2019—2019 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2019年，第6540–6544页。[9] Arsha Nagrani、Joon Son Chung、Samuel Albanie 和 Andrew Zisserman，“利用跨模态自监督学习解耦语音嵌入”，载于*ICASSP 2020—2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2020年，第6829–6833页。[10] Ranya Aloufi、Hamed Haddadi 和 David Boyle，“通过解耦表示保护隐私的语音分析”，载于*2020*年*ACM SIGSAC*云计算安全会议研讨会论文集，2020年，第1–14页。[11] Jarred Barber、Yifeng Fan 和 Tao Zhang，“端到端 Alexa 设备仲裁”，载于*ICASSP 2022—2022 IEEE*声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第926–930页。[12] Joe C Chen、Kung Yao 和 Ralph E Hudson，“声源定位与波束成形”，*IEEE*信号处理杂志，第19卷，第2期，第30–39页，2002年。[13] Fredrik Gustafsson 和 Fredrik Gunnarsson，“利用到达时间差测量进行定位”，载于*2003*年*IEEE*声学、语音与信号处理国际会议论文集（*ICASSP'03*）。IEEE，2003年，第6卷，第VI–553页。

[14] Kenichi Kumatani、John McDonough、Jill Fain Lehman 和 Bhiksha Raj，《基于多通道互相关系数的远场语音识别通道选择》，载于*2011*年免提语音通信与麦克风阵列联合研讨会。IEEE，2011年，第1–6页。[15] Martin Wolf 和 Climent Nadeu，《多麦克风语音识别的通道选择指标》，《style id='5'>《语音通信》，第57卷，第170–180页，2014年。[16] SM Dance 和 BM Shield，《用于预测非扩散封闭空间声场分布的完整镜像声源法》，《声学与振动杂志》，第201卷，第4期，第473–489页，1997年。[17] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin，《注意力机制就是你所需要的一切》，神经信息处理系统进展，第30卷，2017年。[18] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli，《wav2vec 2.0：一种用于语音表征自监督学习的框架》，神经信息处理系统进展，第33卷，第12449–12460页，2020年。[19] Tianyu Gao、Xingcheng Yao 和 Danqi Chen，《SimCSE：句子嵌入的简单对比学习》，载于*2021*年自然语言处理实证方法会议论文集，线上及多米尼加共和国蓬塔卡纳，2021年11月，第6894–6910页，计算语言学协会。