arXiv:2302.05756v1 [eess.AS] 11 Feb 2023

利用自监督学习的语音表征，改进多说话人环境中的注意选择解码

Cong Han[1], Vishal Choudhari[1],Yinghao Aaron Li[1] 和 Nima Mesgarani[1]

摘要——**听觉注意力解码（AAD）是一种用于在嘈杂环境中识别并增强听者所关注说话人声音的技术。其方法是将听者的脑电波与所有声源的表征进行比较，找出最匹配的声源。通常，声音的波形或频谱图可用作声源表征，而这些表征对 AAD 的有效性尚不确定。本研究考察了使用自监督学习的语音表征来提高 AAD 准确率和速度的效果。我们使用侵入式脑电图（ECoG）记录了三名受试者的脑活动；受试者聆听两段对话，并专注于其中一段。我们使用 WavLM 提取每位说话人的潜在表征，并训练时空滤波器，将脑活动映射到语音的中间表征。在评估过程中，将重建的表征与每位说话人的表征进行比较，以确定目标说话人。结果表明，与语音包络和频谱图相比，WavLM 的语音表征能够带来更高的解码准确率和更快的解码速度。研究结果展现了自监督学习的语音表征在听觉注意力解码中的优势，并为开发脑控可穿戴听力技术铺平了道路。**

关键词— **听觉注意力解码、自监督学习、预训练语音模型**

I. 引言

听力障碍听众在嘈杂或拥挤的环境中往往难以理解语音。脑控助听设备可以通过选择性地放大听众正在注意的说话人的语音来提供帮助 [1]。这种设备依赖一种称为注意力听觉解码（AAD）的技术，该技术利用脑活动判断听众正在注意哪位说话人。获取用于注意力听觉解码的脑活动有多种方式，例如脑磁图（MEG） [2],脑电图（EEG） [3], 和皮层脑电图（ECoG） [4]。大多数注意力听觉解码算法会根据脑活动重建语音表征，并将其与环境中的所有说话人进行比较。相关性最高的说话人被视为注意对象说话人。通常，注意力听觉解码所使用的声音表征是包络或频谱图。然而，目前尚不清楚这两者是否都适用于神经解码。用于注意力听觉解码的良好声音表征应当易于从脑活动中重建。尤其是，由于脑信号记录存在局限性，从脑信号到该表征学习复杂的非线性映射可能颇具挑战。此外，良好的表征与注意对象说话人的相关性应高于与非注意说话人的相关性。

[1]哥伦比亚大学电气工程系，纽约州纽约市。Cong Han：ch3212@columbia.edu；Nima Meagarani：nima@ee.columbia.edu

语音自监督表征学习（SSL）已成功应用于许多领域 [5], [6],[7]。SSL通过设计的前置任务学习表征，其中输入和学习目标均取自输入信号本身。因此，SSL可以利用大量无标注语音数据轻松扩展。自监督学习得到的表征通常用作下游任务的输入特征，以减少对大量有标注训练数据的需求并提升任务表现。研究表明，学习得到的语音表征能够改善多种下游任务，例如语音识别、说话人识别和意图分类 [8]。近期一项研究指出，自监督语音模型潜在层的功能层级与语音处理的皮层层级高度吻合[9]。此外，与人工设计的声学特征相比，学习得到的表征在预测语音皮层响应方面更为有效 [10], ，这促使人们提出一种设想：这些表征可能与注意力有关，并且优于注意力听觉解码任务中使用的传统声学特征。

在本研究中，我们使用自监督学习语音表征来改进对注意选择的神经解码。我们使用预训练 WavLM 模型的中间层 [7] 作为脑信号的重建目标，而不是采用传统的语音包络和频谱图特征。实验结果表明，使用这些学习得到的语音表征时，解码准确率显著提高。此外，我们将 WavLM 模型修改为因果配置，以测试其实时实现的潜力；结果发现，它仍优于语音包络和频谱图特征，这表明基于 Transformer 的自监督表征是脑控可穿戴听力设备的更优候选方案。

II. 相关工作

A. 注意力听觉解码的刺激重建

神经科学的最新研究表明，从人类听觉皮层神经反应中重建的频谱图，与受试者专注聆听的说话人的频谱图最为相似 [4]。为了确定注意对象说话人，研究人员会将从神经信号中重建的频谱图与各个说话人的频谱图进行比较。通常使用线性方法进行这种映射，即对电极组应用线性时不变滤波器，以预测注意对象语音的频谱图。通过最小化重建频谱图与实际频谱图之间的均方误差来优化线性滤波器。语音包络

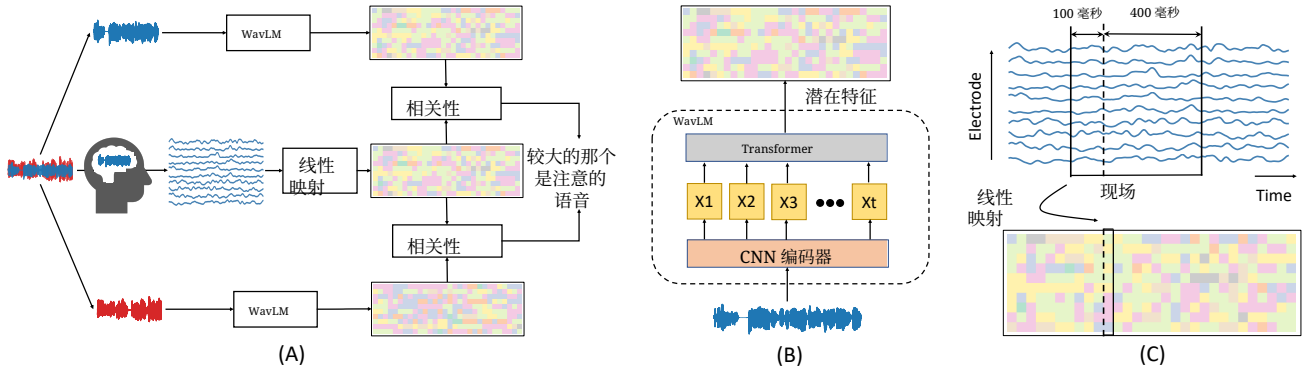

图 1. 听觉注意力解码过程示意图。(A) 受试者听取两位说话人的混合语音并专注于其中一位时，对其神经活动进行监测。WavLM 通过语音分离模型提取单个说话人的表征。将根据神经活动预测的表征与各个说话人的表征进行比较，以确定最相似的说话人。(B) WavLM 由 CNN 编码器和 Transformer 组成，可生成多层特征。其中一个中间层被用作语音表征。(C) 线性时空滤波器将时间延迟范围从 −400 毫秒到 100 毫秒的神经活动映射到学习得到的表征。

也可用作重建目标，尤其是在神经数据的信噪比较低时，例如脑电图数据的情况。本研究旨在探讨，从神经活动中重建自监督学习的语音表征，能否提高解码注意对象说话人的准确率。

B. WavLM

近年来，人们提出了多种自监督语音表征学习方法，其中 wav2vec 2.0 [5] 和 HuBERT [6] 是最知名的两种方法。两个模型的架构相似，但预训练策略不同。wav2vec 2.0 使用对比损失区分正样本和负样本，而 HuBERT 采用离线聚类方法为语音单元分配标签，随后通过类似 BERT 的掩码语音预测任务训练模型。这促使模型从无标注语音数据中同时学习声学特征和语言特征。作为 HuBERT 的变体，WavLM 在预训练过程中加入了语音去噪任务，从而提升其处理非 ASR 任务（如语音说话人分离和语音分离）的能力。在 9.4 万小时多样化语音数据上训练的 WavLM Large，在 SUPERB [8]，上的表现优于以往的自监督语音模型，展现出对语音、说话人、内容和语义进行建模的强大能力。鉴于其出色的性能，我们在本研究中选择了 WavLM Large，因为它最有潜力提升注意或非注意说话人分类的准确率。因此，注意选择解码被视为 WavLM 的下游任务。

III. 材料与方法

A. 神经数据采集与预处理

我们采用侵入式颅内脑电图（iEEG），测量了三名正在接受癫痫治疗的神经外科患者的神经活动。两名患者（受试者 1 和 2）在大脑左半球植入了立体定向脑电图（sEEG）深部电极以及硬膜下脑电图（ECoG）网格电极。第三名患者（受试者 3）仅植入了 sEEG 深部电极植入在其大脑左半球。所有受试者均自述听力正常，并同意参与实验。

每位受试者参与了28个试次。在每个试次中，受试者聆听两段对话的混合音频，同时伴有背景噪声。每段对话均由两名以英语为母语的美国说话者轮流进行。这两段对话同时进行、彼此独立，并从受试者水平面的前半部方向传来。两段对话的功率相同。背景噪声为街道噪声或多人谈话噪声。实验提示受试者将注意力集中在其中一段对话上，并忽略另一段。为确保受试者投入任务，我们在两段对话流中都插入了重复词（RWs），并要求受试者每当听到所跟随对话中的重复词时按下按钮。受试者在所提示的对话中检测重复词的准确率如下：S1——73.2%，S2——74.8%，S3——67.3%。

我们按照[1]中“数据预处理与硬件”部分所述的相同步骤，对神经数据进行预处理，以提取其高伽马（70 - 150 Hz）包络。如果电极对语音的响应显著高于对静音的响应，则选用这些电极。受试者S1、S2和S3分别有34、42和17个语音响应电极。神经数据的采样率为100Hz。

B. 语音表征的提取

WavLM Large 模型 [1] 由一个卷积编码器和 24 个 Transformer 层组成。卷积编码器将以 16 kHz 采样的波形转换为帧率为 50 Hz 的特征序列（每 20 ms 一帧），每帧编码约 25 ms 的波形。每个 Transformer 层的嵌入维度为 1024，并包含 12 个自注意力头。

我们使用 WavLM 从第 i 个层提取语音波形的潜在表征 $X \in \mathbb{R}^{1024*T}$，其中 T 是时间帧的数量。我们将 X 上采样至 100 Hz，以

[1]预训练模型可在 https://github.com/microsoft/unilm/tree/master/wavlm 获取

匹配神经数据的采样率。由于语音持续时间较长，我们将 Transformer 层中每个帧的注意力范围限制为 6 秒（～ 300 个时间帧），即该帧之前 3 秒和之后 3 秒。然而，WavLM 是非因果模型，因为每个时间帧都会关注未来的时间帧。为了与语音包络和频谱图进行公平比较，并实现实时注意力听觉解码（AAD），我们将 WavLM 修改为因果配置。首先，我们将所有未来帧的注意力权重设为零，以强制每个帧只能关注过去 6 秒的内容，并将该模型称为“带因果注意力的 WavLM”。WavLM 中的 Transformer 层配备了基于卷积的位置嵌入，其中卷积操作可以访问未来帧，从而导致非因果计算。为避免这种情况，我们将非因果卷积改为因果卷积，并将所得模型称为“带因果注意力和位置编码的 WavLM”。请注意，在修改注意力权重或位置嵌入后，我们没有对 WavLM 进行微调。考虑到可能导致 WavLM 性能提升的随机效应，我们加入了随机初始化的 WavLM 作为对照。

C. 用于注意力解码的表征重建

我们采用了线性重建方法 [11]。具体而言，学习了两个针对每位受试者的线性时空滤波器G_A 和G_u ，用于将神经活动 R 映射到受注意说话人（$\hat{X}_A$）和未受注意说话人（$\hat{X}_U$）的语音表征，

$$\hat{X}_A(n,t) = \sum_e \sum_\tau G_A(n,e,\tau) R(e,t-\tau) \quad (1)$$

$$\hat{X}_U(n,t) = \sum_e \sum_\tau G_U(n,e,\tau) R(e,t-\tau), \quad (2)$$

其中，n 为表征的通道索引，e 为神经电极索引，τ 为时间滞后；在本研究中，其范围为–400 毫秒至 100 毫秒。与之前一样，通过最小化重建表征与实际表征之间的均方误差来优化线性滤波器。

采用留一交叉验证方法：使用 N - 1 个试次训练针对每位受试者的滤波器，并在留出的试次上重建表征 $\hat{X}_A, \hat{X}_U$。我们计算重建表征 $\hat{X}_A, \hat{X}_U$ 与两位说话人的表征X_{sp1}, X_{sp2}之间的皮尔逊相关系数。相关系数是在一个以秒为单位的窗口内估算的，该窗口称为解码窗口时长。在整个试次时长内，我们分别使用了 0.5 秒、1 秒、2 秒、4 秒和 8 秒的滑动窗口。我们定义了一个注意调制指数（平均互信息），

$$AMI = corr(\hat{X}_A, X_{sp1}) - corr(\hat{X}_A, X_{sp2}) + corr(\hat{X}_U, X_{sp2}) - corr(\hat{X}_U, X_{sp1}). \quad (3)$$

该指标为正值，表明说话人1是注意对象说话人；为负值则表示在此时间窗内，说话人2被判定为注意对象说话人。解码准确率定义为正确分类的时间窗所占的百分比。

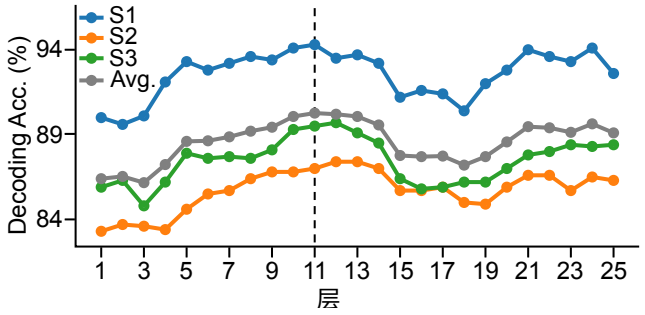

图 2. 使用 WavLM 各层表征进行注意力解码的准确率，解码窗口为 4 秒。第 11 层在各受试者中的平均表现最佳。

IV. 结果

A. *WavLM* 各层的解码准确率

图2展示了 WavLM 不同层的表征对解码准确率的影响。随着层深增加，准确率逐渐提高，之后略有下降，随后再次上升。对于三名受试者而言，第11层的平均表现最佳。第一层是 CNN 编码器的输出，可从语音中提取局部特征（～ 25 毫秒），类似于频谱图。后续各层包含语义信息，且语境更加丰富。近期一项针对 wav2vec 2.0 的逐层分析发现，逐层表征的演变呈现声学-语言层级：浅层编码局部声学信息，随后依次编码语音学信息、词语身份和词义 [12]。因此，图2表明，从大脑解码语音的高层特征，或许能提高注意力解码准确率。 [12]也注意到，从中间层开始出现相反的趋势，并将其归因于 Transformer 层的自编码器式行为：越深层的表征越接近输入。

B. 不同重建目标的解码准确率

表 I 比较了用于听觉注意力解码的不同特征的结果。声学特征包络和 28 基梅尔频谱图是基线特征。包络和梅尔频谱图特征在训练和推理前均进行了 Z 分数标准化。结果表明，从 WavLM 中提取的所有特征始终优于基线（配对t检验，窗口大小为 0.5 s、1 s、2 s 和 4

表 I 使用从 CLEANspeech 中提取的各种特征进行注意力解码的准确率（3 名受试者的平均值，单位：%）

特征	解码窗口大小				
	0.5s	1s	2s	4s	8s
包络	63.3	71.6	79.5	86.0	91.3
梅尔频谱图	65.6	72.3	80.8	88.5	91.5
WavLM	**72.9**	**78.7**	**85.2**	**90.3**	**92.6**
WavLM（因果注意力）	72.2	78.5	84.6	89.4	92.3
WavLM（因果注意力和位置编码）	72.0	77.9	84.1	89.1	92.5
WavLM（随机初始化）	62.8	68.4	74.1	79.8	87.1

表 II 使用从分离语音中提取的各种特征进行注意力解码的准确率（3 名受试者的平均值，单位：%）

特征	解码窗口大小				
	0.5s	1 秒	2 秒	4 秒	8s
包络	63.0	70.1	78.5	85.2	90.1
梅尔频谱图	64.6	71.1	79.2	86.3	90.4
WavLM	**72.2**	**77.8**	**83.9**	**88.6**	**92.4**
WavLM（使用因果注意力）	70.8	76.6	82.5	88.1	92.1
WavLM（使用因果注意力和位置编码）	70.6	76.1	82.3	88.1	91.7
WavLM（随机初始化）	62.8	68.4	74.1	79.8	84.6

s 时 $p < 0.001$；窗口大小为 8 s 时 $p < 0.05$）。与基线相比，WavLM 在解码窗口大小较小时表现尤为出色。因果配置导致性能略有下降，但预计进一步微调可以减轻这种下降。使用随机初始化的 WavLM 进行的对照实验结果显著差于基线，这证实了 WavLM 的性能提升源于通过自监督学习获得的更优表示，而非其架构或特征维度。

由于通常无法获得各个说话者的干净语音，我们使用自动语音分离模型[13] 来分离混合语音。表 II 的结果显示，与表 1 相比，准确率略有下降，这是由于语音分离并不完美，但差异很小，且所有特征受到的影响相似。尽管如此，WavLM 仍优于语音包络和梅尔频谱图。

C. 主成分数量

我们使用主成分分析（PCA）降低了 WavLM 特征的维度。表 III 展示了采用不同 PCA 组件数量的 WavLM 特征的解码准确率。尽管组件数量减少时准确率略有下降，WavLM 仍显著优于梅尔频谱图。

D. 注意力动态切换

我们通过拼接每个试次中神经反应的前 10 秒和最后 10 秒，模拟注意力动态切换；在这些试次中，受试者前 10 秒注意 Spk1，之后切换到 Spk2。我们计算了采用 WavLM（因果注意力与 PE）的 AMI 分数，并

表 III 使用具有不同组件数量的 WavLM 特征进行注意力解码的准确率。

特征	解码窗口大小				
	0.5s	1s	2s	4s	8s
梅尔频谱图（28 维）	64.6	71.1	79.2	86.3	90.4
因果 WavLM（1024 维）	70.6	76.1	82.3	88.1	91.7
200 个主成分	70.3	75.7	82.2	88.1	91.7
100 个主成分	70.0	75.5	81.9	87.8	91.7
50 个主成分	69.5	75.2	81.8	87.4	91.4
28 个主成分	69.0	74.8	81.5	87.7	91.7

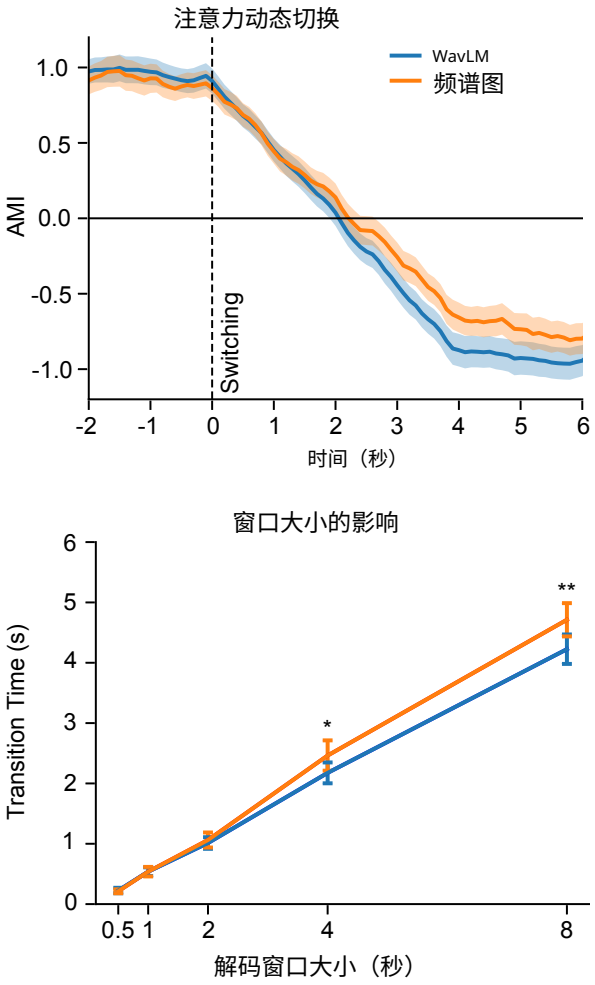

图3。（1）上图显示了从说话人1切换到说话人2时的注意力切换。虚线表示注意力切换发生的时刻。图中显示了三名受试者在4秒解码窗口下的平均 AMI。（2）下图显示了检测到切换的转变时间，该时间以 AMI 穿过零点的时刻进行测量。统计显著性用星号表示：* 表示 $p < 0.05$，** 表示 $p < 0.01$。

梅尔频谱图，分别采用4秒的滑动窗口。图3上方面板显示了所有受试者和试次的平均 AMI 分数。平均 AMI 分数缩放至 -1 到 1 之间。WavLM 和频谱图呈现出相似的模式，但 WavLM 检测到切换的速度更快。图3下方面板显示了五种不同滑动窗口时长下的平均转变时间。与预期一致，窗口时长越长，转变时间越长。窗口大小为2秒及以下时，WavLM 与频谱图之间没有显著差异（配对t检验，$p > 0.2$）；但是，窗口大小为4秒（$p < 0.05$）和8秒（$p < 0.01$）时，WavLM 的转变时间更短。

E. 特征在预测神经活动中的比较

为进一步了解自监督学习特征为何能提高神经解码准确率，我们使用前向模型，根据 WavLM 不同层的特征预测单个神经位点的响应。刺激重建方法使用反向模型，而在此处，我们训练了前向模型，

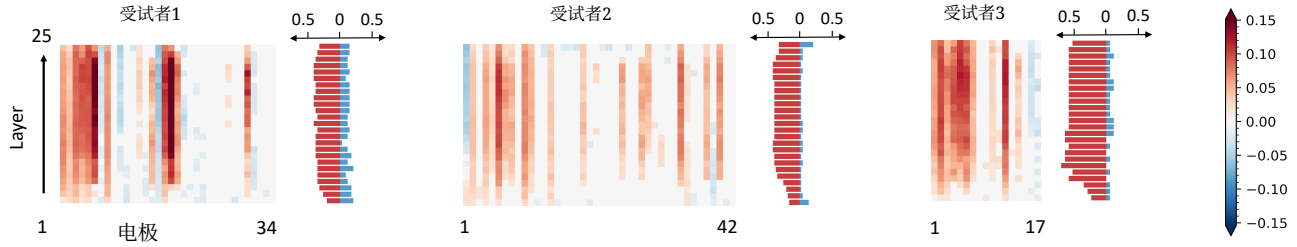

图4. 对每一层和每个电极，使用 WavLM 特征与使用频谱图相比，实际神经活动与预测神经活动之间 r 值的提升。正值表示 WavLM 特征对该电极神经活动的预测效果优于频谱图，负值则表示频谱图的预测更准确。零值（以白色显示）表示两种特征之间没有显著差异（$p > 0.05$）。柱状图分别显示对 WavLM 特征（红色）和频谱图（蓝色）响应更强的电极比例。

时空滤波器，G_R，根据各种刺激特征预测神经活动，

$$\hat{R}(e,t) = \sum_{e}\sum_{\tau} G_R(e,n,\tau)X_A(n,t-\tau), \qquad (4)$$

其中时间滞后 τ 范围为 0 至 200 毫秒。我们测量了每个电极重建的神经活动与实际神经活动之间的相关性（r 值）。我们评估了使用 WavLM 各层相较于梅尔频谱图时 r 值的提升，其中使用了 WavLM 特征的前 100 个主成分。如果配对t检验结果显示，使用 WavLM 特征与梅尔频谱图之间没有统计学差异（$p > 0.05$），则将提升值设为零。此外，我们还计算了每一层中使用频谱图和使用 WavLM 特征能够更准确预测的电极比例。

三个受试者的结果如图 4 所示。与浅层和最深层相比，WavLM 的中间层通常能提供更好的预测。尽管使用声学频谱图（蓝色所示）能够更准确地预测部分电极，但使用 WavLM 特征（红色所示）能够更准确地预测的电极比例更高。图 4 的结果表明，听觉皮层的不同区域编码了不同层次的语音信息。这启发我们结合 WavLM 不同层的特征，以进一步提高听觉注意解码准确率，这将是未来工作的重点。

V. 结论

本研究考察了在多说话人场景中使用自监督语音表征增强注意解码的效果。结果表明，以 WavLM 的潜在特征替代传统语音特征，能够提高注意力解码的准确率和速度，为实现可用的脑控助听设备铺平了道路。这些发现表明，有必要进一步探索自监督语音表征在听觉神经解码中的应用及其潜力，从而增进我们对人脑如何进行注意选择的理解。

致谢

本研究由美国国立卫生研究院（NIH-NIDCD）资助，并获得 Marie-Jos´ee 和 Henry R. Kravis 的资助。我们使用 ChatGPT 提高了文稿的清晰度和可读性。

参考文献

[1] C. Han, J. O'Sullivan, Y. Luo, J. Herrero, A. D. Mehta, and N. Mes- garani, "无需获取干净语音源的说话人无关听觉注意力解码"，《科学进展》，第 5 卷，第 5 期，第 eaav6134 页，2019 年。[2] N. Ding and J. Z. Simon, "聆听竞争说话人时听觉对象神经编码的出现"，《美国国家科学院院刊》，第 109 卷，第 11 期，第 859 页，2012 年，第 29 期，第 854 页。[3] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "鸡尾酒会环境中的注意选择可通过单试次脑电图解码"，《大脑皮层》，第 25 卷，第 1706 页，2015 年，第 7 期，第 1697 页。[4] N. Mesgarani and E. F. Chang, "多说话人语音感知中注意对象说话人的选择性皮层表征"，《自然》，第 485 卷，第 7397 期，第 233–236 页，2012 年。[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0：自监督语音表征学习框架"，神经信息处理系统进展，第 33 卷，第 12 449–12 460 页，2020 年。[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT：通过掩码预测隐藏单元进行自监督语音表征学习"，*IEEE/ACM* 音频、语音与语言处理汇刊，第 29 卷，第 3460 页，2021 年，第 3451 页。[7] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao 等，"WavLM：面向全栈语音处理的大规模自监督预训练"，*IEEE* 信号处理精选主题期刊，第 16 卷，第 6 期，第 1505–1518 页，2022 年。[8] S.-w. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin 等，"SUPERB：语音处理通用性能基准"，载于*Interspeech*，2021 年，第 1194–1198 页。[9] J. Millet, C. Caucheteux, P. Orhan, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, and J.-R. King, "借助自监督学习构建大脑语音处理的现实模型"，*arXiv* 预印本 *arXiv:2206.01685*，2022 年。[10] A. R. Vaidya, S. Jain, and A. Huth, "自监督音频模型能够有效解释人类对语音的皮层反应"，载于国际机器学习会议。PMLR，2022 年，第 21 927–21 944 页。[11] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "语境和行为对初级听觉皮层神经活动刺激重建的影响"，《神经生理学杂志》，第 102 卷，第 6 期，第 3329–3339 页，2009 年。[12] A. Pasad, J.-C. Chou, and K. Livescu, "自监督语音表征模型的逐层分析"，载于 *2021* 年 *IEEE* 自动语音识别与理解研讨会（*ASRU*）。IEEE，2021 年，第 914–921 页。[13] C. Han, Y. Luo, and N. Mesgarani, "保留空间线索的移动说话人双耳语音分离。"载于*Interspeech*，2021 年，第 3505–3509 页。