arXiv:2602.01247v1 [cs.LG] 1 Feb 2026

机制可解释性 脑到语音模型跨语音模式的可解释性

Maryam Maghsoudi * 1 **Ayushi Mishra** * 2

摘要

脑到语音解码模型在发声、默念和想象语音中展现出稳健的性能；然而，这些模型如何在不同语音模态之间捕获并传递信息，其基本机制仍未得到充分探索。在本研究中，我们运用机制可解释性，对神经语音解码器的内部表征进行因果研究。我们在不同语音模式之间对内部激活执行跨模式激活修补，并使用三模态插值考察语音表征是离散变化还是连续变化。我们采用由粗到细的因果追踪和因果擦除来寻找局部化的因果结构，从而发现足以支持跨模式迁移的内部子空间。为确定这些效应在层内的分布精细程度，我们在神经元层面执行激活修补。我们发现，影响跨模式迁移的并非孤立单元，而是规模较小但并非广泛分布的神经元子集。结果表明，各种语音模式位于一个共享的连续因果流形上，而跨模式迁移由紧凑的层特定子空间介导，而非由弥散的活动介导。总体而言，我们的研究结果从因果角度解释了语音模态信息如何在脑到语音解码模型中组织和使用，揭示了跨语音模式的层级化且具有方向依赖性的表征结构。

1. 引言

能够直接从神经活动中解码语音的脑机接口（BCI），有望为无法说话的人恢复自然交流能力。神经解码模型的最新进展展现出令人瞩目的表现，能够根据神经记录重建语音声学特征，尤其是在外显或发声语音方面（Saeidi 等人，2021年）（Livezey 和 Glaser，2021年）（Wu 等人，2024年）（Pescatore 等人，2025年）（Zheng 等人，2025年）。然而，语音脑机接口能否最终应用于实际和临床场景，取决于其能否在没有发声或发声有限的情况下解码语音，例如默念或纯粹的想象语音；这类情形颇具挑战性，因为其时间对齐存在变异（Maghsoudi 等人，2025年）。

尽管实证成果日益显著，但我们对神经语音解码模型如何在内部表征和处理不同语音模式，仍存在一个根本性的认知缺口。针对发声、默念和想象语音训练的模型，其性能差异很大，但尚不清楚这些差异究竟源于神经输入本身不可约的特性，还是源于模型内部表征的形成与利用方式。若缺乏对这些表征的机制性解释，跨语音模式的迁移学习、鲁棒性和泛化能力的进展仍将主要依赖经验，且缺乏系统性。

机器学习社区在机制可解释性方面取得了迅速进展。这一框架通过对内部激活进行因果干预，而非依赖事后相关性，来理解神经网络（Sujatha Ravin-dran & Contreras-Vidal，2023年）。激活修补、因果追踪和子空间干预等技术，已能在语言建模和视觉等领域，将模型行为精确归因于内部表征（Goldowsky-Dill 等人，2023年；Zhang& Nanda，2023年）。然而，这些工具很少被应用于脑到语音解码模型（Sharkey 等人，2025年），而理解其内部机制的重要性，或许不亚于提升原始性能。

本研究采用机制可解释性方法，并通过受控干预探究网络如何表征和传递有关各种语音模式的信息。神经语音解码器使用三种类型的语音进行训练：发声、默念和想象。在保持输入不变的同时，将内部激活从一种模式移至另一种模式，我们由此检验准确解码所需的表征。本研究还探讨这些表征是否

*共同贡献 [1]马里兰大学帕克分校电气与计算机工程系，美国 [2]马里兰大学帕克分校计算机科学系，美国。通讯作者：Maryam Maghsoudi <maryam00@umd.edu>，Ayushi Mishra<amishr13@umd.edu>。

预印本。*2026*年*2*月*3*日。

会在不同语音模式间以清晰的阶段变化，还是逐渐变化；以及重要信息是分散分布，还是集中在网络的特定部分。

我们的结果表明，语音模式之间的差异编码在共享的、分级的内部表征中。高性能的发声表征可以因果地迁移到想象输入上，从而恢复解码质量；而反向干预则会使性能灾难性下降。我们表明，跨模式迁移由紧凑且结构化的内部子空间介导，这些子空间既体现在循环动态的时间维度上，也体现在卷积特征的空间维度上，而不是由遍布整个网络的弥散性活动所驱动。

这项工作超越了描述性的性能比较，转向对其原理的理解。我们的发现不仅对改进想象语音脑机接口具有意义，也有助于设计具备可解释性、鲁棒性并能够实现可靠跨模式泛化的模型。

2. 相关工作

近年来，可解释人工智能（XAI）的进展使得深度学习系统和大语言模型（LLM）能够应用于许多临床决策场景（Kim等，2024年）（Mishra等）。在这些领域，模型需要以可靠且透明的方式解读复杂的生物医学信号（Shimizu等，2025b）（Fucci等，2024年）（Norval和Wang，2025年）（Bibi等，2025年）（Bibi等，2025年）（Maurer等，2024年）。理解这些模型如何做出预测，对于确保安全性与可信度至关重要（Jain等，2021年）（Davies和Khakzar，2024年）（Plantinga等，2025年）。语音脑机接口（BCI）正是这一挑战的一个实例，因为它们利用神经网络从脑记录中解码语音信息。DeWave（Duan等，2023年）探索了基于脑电图的语音与语言解码，包括脑电图到文本的转换和无标记解码，但主要强调终端任务性能。这些方法并未考察跨模式性能差异如何在训练后的解码器内部产生，也未探究这些差异究竟源于输入信号的局限，还是源于模型内部表征。

机制可解释性尝试通过改变内部激活，逆向工程学习到的表征与计算，而非仅使用相关性探针（Kowalska 与 Kwa´snicka，2025年）。激活修补也称为因果追踪或交换干预，是其中一种关键方法（Heimersheim 与 Nanda，2024年）。近期的最佳实践建议强调，应谨慎选择基线、扰动分布和指标（Bereska 与 Gavves）（Conmy 等，2023年）（Glazer 等，2025年）（Williams 等，2025年）。电路导向方法更进一步，不局限于单层干预，而是寻找因果组件及其相互作用，通常借助自动化工作流进行电路发现与归因式定位（Cunningham 等，2023年）。多篇综述和立场论文汇总了这些方法，并探讨了开放问题，例如需要更清晰的概念和更可靠的因果干预。在这项工作中，我们将这些因果工具用于神经解码，并重点关注语音模式之间的跨条件迁移，将其作为一种自然的干预方式。

近期研究认为，可解释性应与认知和计算神经科学更加紧密地结合（Mineault 等人，2024年），因为理解内部表征是主要的科学目标，而非次要考量。在脑机接口研究中，研究人员通常会分析电极定位、特征重要性或架构消融（Shimizu 等人，2025a）。这些方法可以揭示信号出现在哪里，但通常无法告诉我们，某些内部表征是否确为解码所必需或是否足以支持解码。通过采用激活层面的干预、因果定位和擦除，我们更清晰地说明了语音模式信息如何在训练后的解码器中表征。我们的方法补充了现有以性能为重点的研究，揭示了在不同语音模式下，支配解码行为的是模型的内部表征，而不仅仅是输入信号。

3. 跨言语模式的机制可解释性

神经语音解码模型在各种语音条件下的表现日益出色，包括发声、默念和想象语音 [**附录A**]。然而，目前仍不清楚这些模型如何在内部表征不同的语音模式。

我们的目标不仅是衡量内部激活与输出行为之间的相关性，还要找出哪些内部表征在因果上决定了不同模式下的语音解码。为此，我们采用受机制可解释性领域近期关于激活修补和因果追踪的研究启发的干预分析方法（Bereska 与 Gavves）。这些方法允许我们在保持输入和其余计算不变的情况下，修改特定的内部激活，从而对内部表征的作用进行反事实推理。

至关重要的是，多种语音模式的存在为因果分析提供了天然的研究环境。发声、默念和想象语音具有相同的高层语言内容，但神经信号特性和解码性能各不相同，因此我们可以探究：在一种模式下学得的内部表征能否替代另一种模式的表征、使其性能下降，或与之进行插值。这种跨模式结构使我们能够开展超越标准消融实验的干预，直接检验不同条件下内部表征的充分性、必要性和连续性。

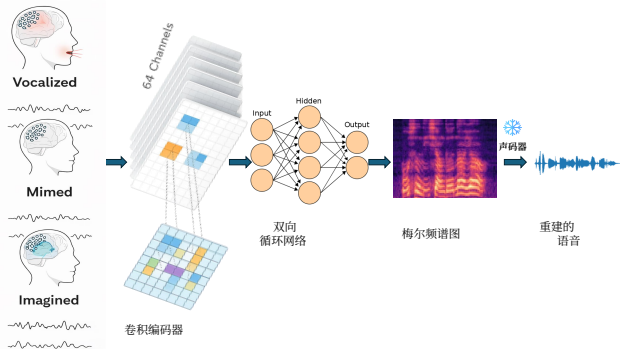


图*1.* 多通道立体脑电信号经卷积编码器处理，以提取时空特征；随后，双向循环网络整合时间信息，并预测帧级梅尔频谱图。神经声码器将预测的频谱图转换为重建语音。

4. 数据集与模型架构

4.1. 发声、默念和想象语音的立体定位脑电数据集

我们在 VocalMind (Heetal., 2025年）数据集上评估我们的方法。该大规模立体脑电图（sEEG）数据集旨在支持多种语音模式下的神经语音解码。数据集包含一名人类参与者在三种条件下执行语音任务时的神经记录：发声、默念和想象语音。**发声语音**（外显语音）是指通过发声产生声音的构音。**模拟语音**是指不产生声音的无声构音。**想象语音**是指不进行外显构音的内部语音生成；本文聚焦于与想象说话相对应的语音运动想象。

与大多数以单一语音模式为重点的既有神经语音数据集不同，VOCALMIND 支持对相关说话条件进行系统比较。这种多模态结构非常适合研究不同语音模式之间的表征迁移和因果机制。

4.2. 模型架构

如图所示，该模型由三个主要部分组成1。首先，卷积编码器从原始立体脑电信号中提取跨通道和时间的局部时空特征。具体而言，该编码器首先使用一维卷积层（`Conv1D`），沿时间维度进行卷积，输出通道数为 64 、卷积核大小为 4 、步幅为 4 、填充为 2，从而在保留局部时间结构的同时进行时间下采样。随后，将得到的卷积激活输入多层双向循环网络，以整合较长时间上下文的信息，并生成帧级声学预测。

循环部分由门控循环单元（GRU）网络组成，每个方向包含 64 个输入特征和 256 个隐藏单元，并排列为三层堆叠层。最后，全连接线性层将循环特征映射到 80 维输出空间，对应梅尔频谱图帧。随后，神经声码器（Kong 等人，2020）根据预测的梅尔频谱图重建语音波形。

令 $f(x)$ 表示完整模型，其中 x 是一个立体脑电信号输入序列。我们将卷积编码器产生的激活（卷积层输出）记为 z^{conv}，将循环网络产生的激活（循环神经网络层输出）记为 z^{rnn}。这些内部表征构成了以下各节所述机制分析的主要干预位点。除非另有说明，循环部分采用多层双向门控循环单元实现。

5. 跨模式激活修补

我们使用跨模式激活修补研究不同语音模式下内部表征的因果作用。令 X_I表示想象语音模式，X_V 表示发声语音模式。对于给定的层 ℓ，令 Z_I^ℓ 和 Z_V^ℓ 表示通过标准前向传播产生的相应隐藏激活。

在修补后的前向传播中，我们将激活 Z_I^ℓ替换为 Z_V^ℓ，同时保持原始输入 X_I 不变，并将修改后的激活传播至网络的其余部分。令 $f_{\text{post-}\ell}$ 表示由层 ℓ之后所有层组成的子网络。修补后的输出如下：

$$\tilde{y} = f_{\text{post-}\ell}(z_v^\ell). \tag{1}$$

我们将此过程应用于卷积层和循环层，以检验两种互补的因果属性：充分性和必要性。通过将表现较好的语音模式（例如发声语音）的激活修补到表现较差的输入（例如想象语音）中，并测量解码性能是否提升，来评估充分性。通过将表现较差的模式的激活修补到表现较好的输入中，并判断性能是否下降，来评估必要性。

这种双向跨模式修补框架能够将性能差异直接进行因果归因，归因于特定网络层中的内部表征。我们的完整机制性解释如图 2 所示。

5.1. 三模态激活插值

跨模式激活修补可以检验因果充分性和必要性是否是离散的，但无法说明内部表征是否会随语音模式逐渐变化。为了探究语音模式是以彼此独立的类别还是沿连续谱编码的，我们对想象（I）、默念（M）和发声（V）语音进行三模态激活插值。

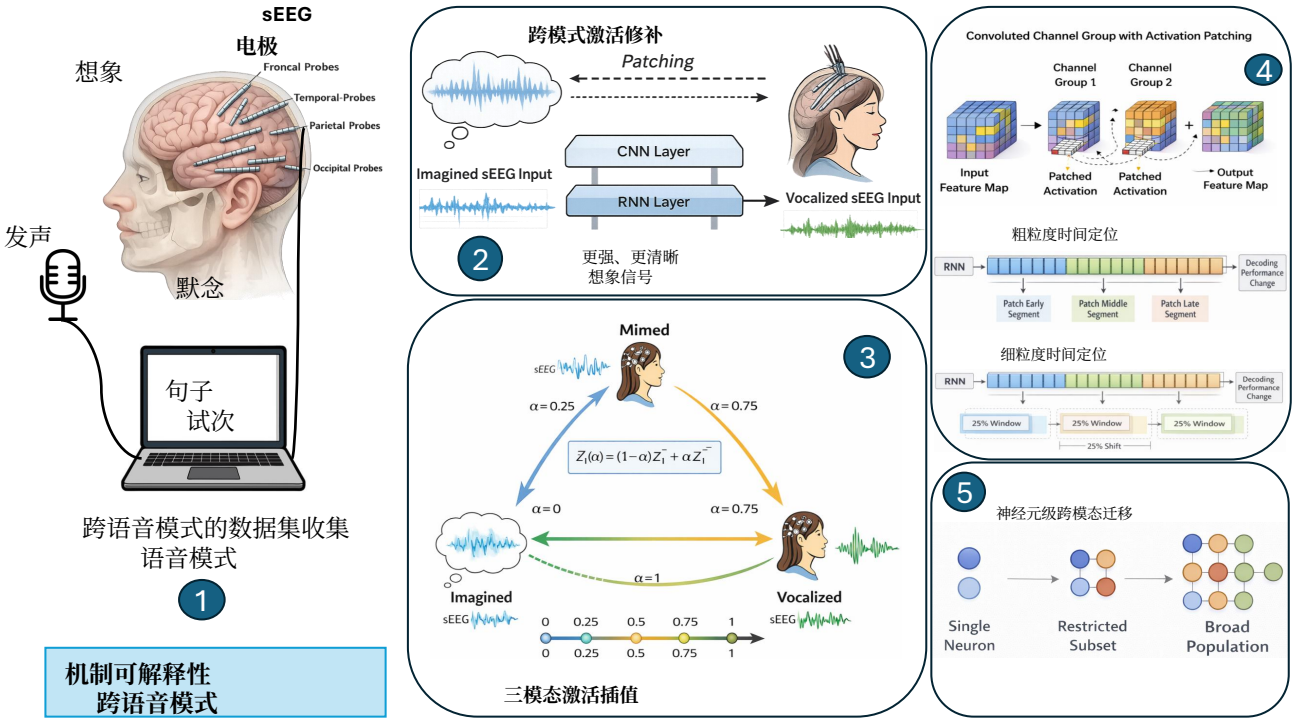


图2。 **跨言语模式的机制可解释性**。 **（1）**我们研究三种模式下的语音解码：发声、默念和想象，并使用立体脑电图记录。 **（2）**跨模式激活修补：我们在选定的层（如卷积神经网络/循环神经网络）上，将一种模式的内部激活替换为另一种模式的内部激活，以检验表征是否能跨模式迁移，以及修补能否产生更强或更清晰的想象信号。 **（3）**三模态激活插值：我们对不同模式间的激活进行线性插值，以探究表征几何和解码行为的平滑变化。 **（4）**时间定位：我们通过修补后续卷积通道，以及循环神经网络激活序列中连续的早期/中期/晚期三等分片段（粗粒度），并以25%的窗口、25%的步长滑动（细粒度），测量由此产生的解码性能变化，从因果层面定位信息何时发挥作用。 **（5）**神经元级跨模态迁移：我们通过修补单个神经元，再修补按其作用分组的前k个神经元，检验迁移是否由单个神经元、受限子集或广泛神经元群体介导。

对于跨模式共享的给定句子，我们首先进行标准前向传播，在目标层 $\ell \in$ {卷积层输出，循环神经网络层输出}提取内部激活，从而得到来自两种语音模式 A 和 B的表征 z_ℓ^A和 z_ℓ^B。然后，我们将激活构造为以下凸组合：

$$\tilde{z}_\ell(\alpha) = (1-\alpha) z_\ell^A + \alpha z_\ell^B, \quad \alpha \in [0,1]. \qquad (2)$$

在一次激活修补前向传播中，插值后的激活$\tilde{z}_\ell(\alpha)$会替换第 ℓ层的原始激活，同时将输入脑电信号固定为 A 模式。在多个插值系数$\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$下评估解码性能。

我们对默念语音和发声语音之间的所有有序配对执行此过程。

这一干预直接检验不同语音模式下的内部表征是否位于共享的因果流形上，以及模式间差异体现的是渐进变化，还是性质迥异的内部状态。

为了进一步理解在网络中何处跨模态信息被实例化，我们接下来进行更详细的因果追踪分析。

5.2. 因果定位与细粒度追踪

在跨模式激活修补与插值的基础上，我们接下来将定位语音相关因果信息在网络中的具体位置。我们通过粗粒度干预识别候选区域，并通过细粒度追踪解析其内部结构。

粗粒度因果定位。 在产生64个特征通道的卷积层，我们将通道划分为四组、每组16个，并一次修补一组，同时保持其余部分固定。该测试检验因果信号是分布在各通道之间，还是集中于特定特征子空间。在循环层，我们将RNN输出序列划分为早期、中期和晚期时间片段，并分别独立修补每个片段，以探究因果信息是否定位于循环处理的某一特定阶段。

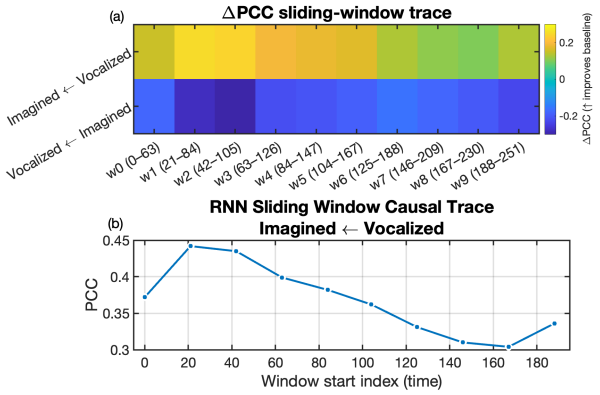


图*3*。 跨语音模式的循环神经网络滑动窗口因果追踪。早期至中期的循环状态构成一个因果瓶颈，对高质量语音解码既充分又必要。

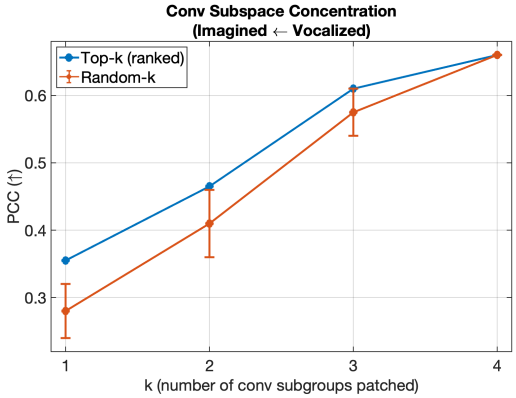


图*4.* 卷积编码器中的因果子空间集中。按排名选取的 Top-k 通道子组修补始终优于匹配随机 -k 对照，这表明跨模式迁移由紧凑的卷积子空间介导，而非由均匀分布的特征介导。

细粒度因果追踪。 我们使用定向干预来细化这些分析。对于循环神经网络，我们通过修补25%的时间窗口，并沿256维序列在10个位置上滑动，应用滑动窗口因果追踪。这会生成随时间变化的因果剖面。如图3所示，将发声循环神经网络状态迁移到想象输入，对早至中期窗口带来最强增益，具体而言，窗口时间步从[21-84],，而较晚窗口的效应较弱；必要性方向在同一区域呈现出互补的退化模式。

我们检验卷积层中的因果影响是否集中在一个低维子空间中。找到效果最佳的粗粒度通道组（g2；通道32至48）后[**附录 D**], 我们将其拆分为更小的组，并根据修补效果对这些组进行排序。图4显示，累积 Top-k 修补带来了稳定的提升，且效果始终优于匹配的随机 k 对照组。这表明，跨模式迁移是通过紧凑的卷积子空间实现的，而非依赖分散的特征。

5.3 局部子空间的因果擦除

我们采用因果擦除来检验通过激活修补和细粒度因果追踪（第5.2节）识别出的内部子空间是否足以实现跨模式迁移，以及这些子空间之外的信息是否可以替换。对于各模式共享的每个刺激键，我们首先计算发声、默念和想象语音的基线解码结果。随后，我们选择一种供体模式，并在解码另一种接收模式时进行干预：在卷积层输出（通道）和循环神经网络层输出（时间）处修补内部激活。

基于由粗到细的因果追踪，我们定义了小型保留区域：每个位置分别包括：**(i)** 卷积通道 $[32, 48)$（16 个通道），以及 **(ii)** 循环时间步 $[21, 84)$（序列的 25%）。

对于每个接收输入，通过在保留区域复制供体激活，并用另一个随机选取的同形状供体样本的激活填充其余部分，构造混合激活。测试了三个版本：**KEEP-Conv**、**KEEP-RNN** 和**KEEP-Combo**。它们分别只保留选定的卷积子空间、循环时间窗口，或两者。

为了控制修补信息量而非子空间身份，我们采用匹配随机对照（**RAND-Conv**、**RAND-RNN**、**RAND-Combo**），保留一个随机选取的同等大小区块。我们还纳入全量修补基线（**fullConv**、**fullRNN**），将整个卷积层输出或循环神经网络层输出替换为供体激活。这给出了迁移可能达到的上限。

5.4. 神经元级激活修补

截至目前的研究都是在网络层这一粗粒度层面进行的。在本节中，我们将转向更精细的尺度，聚焦于层内更小的单元——单个神经元。我们提出两个相关的问题。

首先，跨模态迁移是由单个神经元、有限的神经元子集，还是广泛的群体层面修补所介导？为回答这一问题，我们在不同语音模式间进行单神经元激活修补，并测量逐一修补每个神经元所引起的解码性能变化。随后，我们根据神经元各自的影响对其分组，并将排名前k个神经元一起修补，以检验性能提升或下降是否源于少数神经元的协同作用。

其次，在不同输入样本中，最具影响力的神经元是否一致？为回答这一问题，我们找出每个单独样本中最具影响力的神经元，并分析这些神经元在不同句子中的分布和重复出现情况。

我们首先介绍激活修补的设置，然后展示两个实验来回答上述问题。

设置。考虑一个源输入模式 m_s （修补来源）、一个目标输入模式 m_t （修补目标）、一层 ℓ以及一个神经元索引 i。我们将解码器表示为多个模块的组合，并用 $a_\ell(x)$ 表示模块 ℓ 处理立体脑电图输入 x时产生的激活张量。考虑一对配对输入（$x^{(m_s)}$ $x^{(m_t)}y$），它们具有相同的真实目标 y （梅尔频谱图）。第 ℓ 层的源激活和目标激活分别为 $A_s = a_\ell\left(x^{(m_s)}\right) \in \mathbb{R}^{T_\ell \times D_\ell}$ 和$A_t = a_\ell\left(x^{(m_t)}\right)$。其中， T_ℓ 表示第 ℓ 层的时间步数，而 D_ℓ 表示该层隐藏表示的维度（即神经元数量）

我们将修补后激活 $\tilde{A}_t^{(i)}$ 定义为：在所有时间步中，仅替换第i个神经元：

$$\tilde{A}_t^{(i)}[t,j] = \begin{cases} A_s[t,i] & \text{if } j = i, \\ A_t[t,j] & \text{if } j \neq i, \end{cases} \qquad \forall t \in \{1,\dots,T_\ell\}. \tag{3}$$

我们将目标模态输入送入同一个固定模型，同时在第 ℓ 层进行干预，将神经元 i 替换为源激活轨迹。由此得到的修补后预测为

$$\hat{y}^{(i)} = M_\theta\left(x^{(m_t)};\, a_\ell(x^{(m_t)}) \leftarrow \tilde{A}_t^{(i)}\right). \tag{4}$$

作为比较，我们还计算未修补基线预测 $\hat{}$ $y^{\text{base}} = M_\theta\left(x^{(m_t)}\right)$，并量化相对于 $\hat{}y^{\text{base}}$的提升或退化。此过程将单个神经元的时间激活轨迹对跨模态迁移的因果贡献单独分离出来。

5.5. Top-k 神经元饱和度分析

单神经元修补揭示了各个神经元对解码性能的影响。然而，为了评估性能变化究竟是由协同子集而非孤立单元驱动，我们进行了 top-k 神经元饱和度分析。在此分析中，对于给定的第 ℓ 层和模态对（$m_s \to m_t$），我们首先根据神经元在数据集上的平均单神经元效应 Δ皮尔逊相关系数$_i$对神经元进行排序，得到从因果影响最大到最小的排列 π 。

接下来，我们定义前k个神经元的集合 $S_k = \pi(1)\dots,\pi(k)$，其中包含单独修补后能最大程度提升解码性能的 k 个神经元。我们同时修补 S_k 中的所有神经元，将其激活轨迹从源前向传播替换到目标模型的前向传播中，同时保持所有模型参数不变。第 ℓ 层修补后的激活为

$$\tilde{A}_t^{(S_k)}[t,j] = \begin{cases} A_s[t,j] & j \in S_k, \\ A_t[t,j] & \text{otherwise.} \end{cases} \tag{5}$$

对于所有 $t \in \{1,\dots,T_\ell\}$。

我们评估了解码性能随 k 的变化，得到饱和曲线，其形状反映了因果信息在某一层内各神经元之间的分布方式。

5.6. 跨句子的神经元级效应一致性

为回答第二个问题，我们评估同一批神经元是否负责在不同输入样本上带来改进。为此，我们进行句子级胜者分析。我们固定一个源-目标模态对（$m_s \to m_t$）和一层 ℓ，并针对每个输入样本，将源模态中的所有单个神经元修补到目标模型中。对于每个神经元和每个输入样本，我们测量相对于未修补基线的 Δ皮尔逊相关系数。随后，对于每个输入样本，我们选择 胜者神经元，即在 Δ皮尔逊相关系数方面带来最大改进的神经元。这样，每个句子都会得到一个获胜神经元索引。

随后，我们分析获胜神经元在输入样本中的分布。(i) 我们计算唯一胜者神经元的数量。如果该数量较小，说明只有少数神经元负责在所有输入样本上带来改进；如果该数量较大，则表明这一责任分散在许多神经元之间。(ii) 我们计算出现频率最高的胜者神经元所影响的输入样本数量。(iii) 我们计算出现频率最高的前 k 个获胜神经元覆盖了多少输入样本。如果这前 k 个神经元覆盖了很大一部分样本，则说明同一批神经元往往推动整个数据集上的改进；否则，负责改进的神经元会因句子而异。最后，(iv) 我们计算胜者神经元分布的熵。熵越高，意味着神经元分布越均匀，并参与特定句子。总体而言，这些指标表明，跨模态改进是由一组一致的神经元造成的，还是由多样且针对具体输入的神经元子集造成的。

6. 评估

指标。 我们通过比较修补后预测与未修补基线，使用两种互补指标评估因果效应。我们使用预测梅尔频谱图与真实梅尔频谱图之间的皮尔逊相关系数（PCC）衡量全局重建保真度，并使用梅尔倒谱失真（MCD）衡量频谱准确度。

方向	完整补丁	KEEP-Conv	RAND-Conv	KEEP-RNN	RAND-RNN
PCC					
想象 ←发声（充分性）	**0.954**	**0.666**	0.564	**0.462**	0.398
发声 ← 想象（必要性）	**0.177**	0.575	0.575	0.456	0.460
MCD					
想象 ←发声（充分性）	**1.63**	**3.13**	3.23	3.10	3.26
发声 ← 想象（必要性）	**3.70**	2.95	2.98	**2.88**	3.15

表1. 因果擦除结果仅针对方向性干预。我们将全激活修补与子空间保持擦除（KEEP）以及匹配随机对照（RAND）进行比较。结果针对充分性（想象 ← 发声）和必要性（发声 ← 想象）进行报告。

PCC 通过展平后的频谱图之间的相关性计算得出；MCD 则根据倒谱系数计算，这些系数通过先进行对数变换、再沿频率维度进行离散余弦变换获得。对于这两项指标，我们报告相对于基线预测的变化，其中 Δ PCC 越高、MCD 越低，表示性能越好。

6.1. 跨语音模式的共享因果流形

在各语音模式中，因果干预揭示的是共享的内部结构，而非离散表示。完整的跨模式激活修补表明，发声内部状态足以在想象与默语输入上恢复高质量解码，而用想象状态替换发声状态则会导致性能崩溃。跨模式激活修补的基线结果见**附录B.1**。三模态激活插值进一步表明，在想象、默念与发声语音之间，解码质量呈现平滑且单调的变化，各方向上的排序一致且具有对称性。三种语音模式的性能均随插值系数 α平滑变化。三模态插值实验结果见**附录C**。综合来看，这些结果表明语音模式位于共享的连续因果流形上，模拟语音占据中间表征位置，而非形成独立机制。

6.2. 因果擦除证实集中且定向的迁移

因果擦除用于检验激活修补所识别的子空间是否足以实现跨模式迁移。对于想象 ← 到发声，保留所识别的卷积子空间，其效果显著优于匹配的随机对照；保留 RNN 时间窗口还能带来额外但较弱的提升。这表明，迁移集中在少数内部特征中。表1总结了因果擦除分析的结果。反向迁移中，完整修补会使性能降至想象语音水平，而子空间保持擦除只能带来有限且因指标而异的益处，反映出发声语音与想象语音之间表征完整性的固有不对称性。

性能，而子空间保持擦除带来的益处有限且因指标而异，这反映出发声语音与想象语音之间表征完整性的固有不对称性。

6.3. 神经元级饱和分析

我们观察到，修补单个神经元不会导致皮尔逊相关系数发生较大变化，这表明不存在仅凭自身激活就能驱动跨模态性能的单个神经元。受此观察启发，我们将分析扩展到对神经元组进行修补。具体而言，首先逐个修补神经元，根据其对皮尔逊相关系数的单神经元效应进行排序，然后按照排名以 top-k 的方式联合修补。

图5 展示了皮尔逊相关系数随 k变化的情况。每个子图对应一种目标模态（修补到的模态），曲线则显示从另外两种源模态（修补来源）进行修补的结果。在所有子图中，我们都观察到一个共同模式：随着 Δ增加，皮尔逊相关系数先提升，随后退化。这种饱和—退化曲线表明，跨模态迁移并非源于内部表征的全面替换，而是依赖于一个受限的因果子空间；对该子空间不加区分地扩展会引入干扰。

这些效应的幅度和方向在很大程度上取决于源-目标模态配对。将发声激活修补到默念或想象目标中，可获得最大的皮尔逊相关系数提升（图 5中的绿色曲线）；而修补到发声目标中几乎没有提升，甚至毫无提升，这表明解码器对于发声语音而言已接近最优状态；更多有关模态特异性不对称性的分析见**附录E**。

最后，不同条件下也呈现出架构差异。在修补卷积层神经元能够提升皮尔逊相关系数的情况下（尤其是目标为默念或想象时），峰值性能对应的 k 值高于RNN层中的情况。这种延迟峰值表明，卷积层支持更分布式的因果贡献，而循环层则依赖更小但更精确的

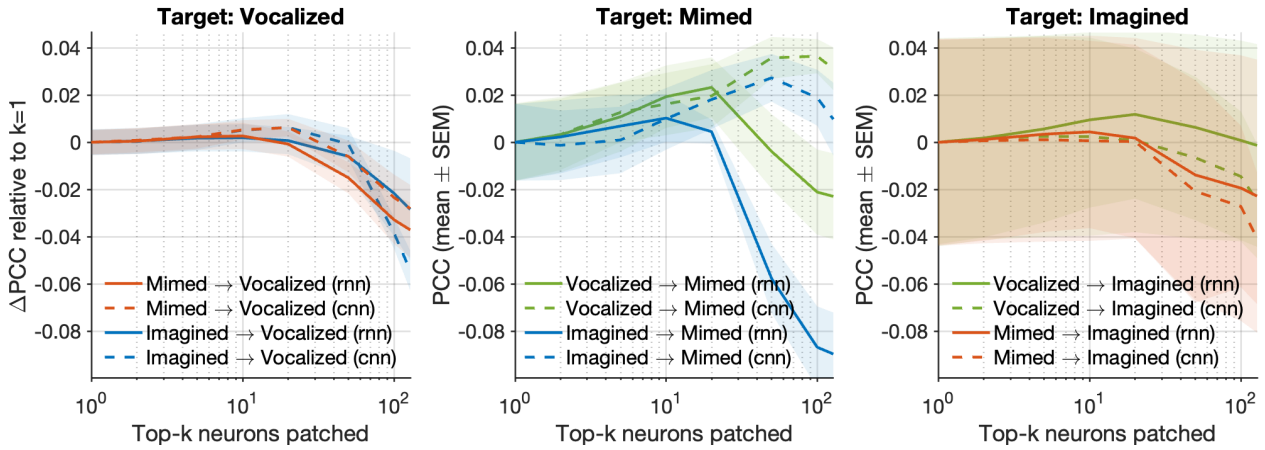


图5. Top-k 神经元修补揭示了语音模态之间不对称的子空间迁移。 ΔPCC（相对于 k = 1）随修补神经元数量的变化情况，针对三种目标模态分别展示：发声（左）、默念（中）和想象（右）。实线表示 RNN 层，虚线表示卷积层。阴影区域表示不同折之间的 ± 标准误。

神经元子集。

6.4. 神经元层面因果效应的句子级结构

为评估神经元层面的因果效应如何跨句子泛化，我们分析了句子级获胜者覆盖率随修补神经元数量（k）的变化情况。覆盖率随着 k 增加，并且在所有源-目标模态对中均如此，这表明 PCC 提升源于多个神经元，而非单个占主导地位的单元。详细的覆盖率曲线和定量汇总见**附录 F**

7. 讨论与未来工作

我们的结果表明，发声、默念和想象语音之间的性能差距，主要源于解码器组织信息的方式，而非输入信号质量的差异。插值结果显示，这三种语音模式处于同一连续体上，其中默念语音始终居于中间。因果干预表明，只有发声语音内部状态才能对想象输入进行高质量解码，并且要维持发声语音水平的性能也离不开这些状态。因果定位发现，跨模式迁移发生在一些狭窄的瓶颈中，例如有限的卷积子空间和循环轨迹中的一小段。

我们的神经元层级因果分析表明，跨模态语音解码依赖于共享但分布不均的神经子空间。此外，发声表征为改善默念和想象语音解码提供了更有效的子空间，这与粗粒度的激活修补结果一致。未来的研究可以将神经元层级的因果结构与立体脑电图中的解剖与功能组织联系起来，从而更深入地了解想象语音和隐性语音的神经基础。

未来工作。 未来研究应检验，在使用其他序列模型（如长短期记忆网络、Transformer 或状态空间模型），以及纳入更多受试者和数据集时，所识别出的卷积和循环瓶颈是否依然存在。观察这些已识别子空间在训练过程中的发展，也有助于解释模型结构如何影响学习。训练还应促使想象和默语输入使用与准确发声解码相同的内部子空间。这样可以提升模型的可解释性，并帮助它们更好地泛化到不同的语音模式。

8. 结论

本研究表明，机制可解释性能够揭示因果性地支持语音解码的内部表征，并阐明这些表征在发声、默念和想象语音之间有何差异。通过将性能差异与局部且可迁移的内部结构联系起来，我们的发现指向这样的神经解码模型：不仅准确度更高，而且可预测性、可控性和鲁棒性也更强。

影响声明

这项研究通过因果干预，探究哪些内部表征会影响发声、默念和想象语音的脑到语音解码表现，从而加深我们对脑到语音解码机制的理解。这些发现有助于通过审慎地支持跨模式泛化，而非依赖反复试错，构建更可靠的辅助通信系统。然而，神经数据高度敏感，因此解码方面的任何进展都必须采取严格的隐私保护措施、取得知情同意并进行清晰沟通，以免夸大模型对人类认知所揭示的内容。

参考文献

Bereska, L. and Gavves, E. 人工智能安全的机制可解释性——综述，2024年。网址 *https://arxiv. org/abs/2404.14082*。

Bibi, N., Courtney, J., and McGuinness, K. 利用可解释人工智能增强脑疾病诊断：近期研究综述。*ACM* 医疗保健计算汇刊，6(2)：1–35，2025年。

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. 面向机制可解释性的自动化电路发现。神经信息处理系统进展，36：16318–16352，2023年。

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. 稀疏自编码器在语言模型中发现高度可解释的特征。*arXiv* 预印本 *arXiv:2309.08600*，2023年。

Davies, A. and Khakzar, A. 可解释性中的认知革命：从解释行为到解释表征与算法。*arXiv* 预印本 *arXiv:2408.05859*，2024年。

Duan, Y.、Zhou, J.、Wang, Z.、Wang, Y.-K. 和 Lin, C.-T. DeWave：用于将脑电图动态编码为文本翻译的离散脑电波。*arXiv* 预印本 *arXiv:2309.14030*，2023年。

Fucci, D.、Savoldi, B.、Gaido, M.、Negri, M.、Cettolo, M. 和 Bentivogli, L. 语音模型的可解释性：关于声学特征选择的挑战。载于第十届意大利计算语言学会议（*CLiC-it 2024*）论文集，第 373–381 页，2024年。

Glazer, N.、Segal-Feldman, Y.、Segev, H.、Shamsian, A.、Buchnick, A.、Hetz, G.、Fetaya, E.、Keshet, J. 和 Navon, A. 超越转录：自动语音识别中的机制可解释性。*arXiv* 预印本 *arXiv:2508.15882*，2025年。

Goldowsky-Dill,N., MacLeod, C., Sato, L., and Arora, A. 使用路径修补定位模型行为。*arXiv*预印本 *arXiv:2304.05969*，2023年。

He, T., Wei, M., Wang, R., Wang, R., Du, S., Cai, S., Tao, W., and Li, H. VOCALMIND：一个用于声调语言中发声、默念和想象语音的立体定位脑电数据集。《科学数据》，12(1):657，2025年。doi: 10.1038/s41597-025-04741-2。

Heimersheim, S. 和 Nanda, N. 如何使用和解读激活修补。*arXiv* 预印本 *arXiv:2404.15255*，2024年。

Jain, S.、Vo, V. A.、Mahto, S.、LeBel, A.、Turek, J. S. 和 Huth, A. G. 用于预测连续自然语音功能性磁共振成像响应的可解释多时间尺度模型。bioRxiv。2021年。Kim, J. W.、Alaa, A. 和 Bernardo, D. EEG-GPT：探索大语言模型在脑电图分类与解读方面的能力。*arXiv* 预印本 *arXiv:2401.18006*，2024年。

Kong, J., Kim, J., and Bae, J. HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。神经信息处理系统进展，33:17022–17033，2020年。

Kowalska, B. and Kwa´snicka, H. 揭开黑箱：用于理解神经网络算法的机制可解释性。*arXiv*预印本 *arXiv:2511.19265*，2025年。

Livezey, J. A. 和 Glaser, J. I. 面向不同架构和记录模态的神经解码深度学习方法。《生物信息学简报》，22(2):1577–1591，2021年。

Maghsoudi, M.、Rezaeizadeh, M. 和 Shamma, S. 一种用于将想象的听觉脑信号映射为聆听时脑响应的卷积框架。*arXiv*预印本 *arXiv:2512.03458*，2025年。

Maurer, M. C.、Metsch, J. M.、Hempel, P.、Bender, T.、Spicher, N. 和 Hauschild, A.-C. 用于临床决策支持的生物信号可解释人工智能。载于第*30*届 *ACM SIGKDD*知识发现与数据挖掘会议论文集，第6597–6604页，2024年。

Mineault, P.、Zanichelli, N.、Peng, J. Z.、Arkhipov, A.、Bing-ham, E.、Jara-Ettinger, J.、Mackevicius, E.、Marblestone, A.、Mattar, M.、Payne, A. 等。面向人工智能安全的神经人工智能。*arXiv*预印本 *arXiv:2411.18526*，2024年。

Mishra, A.、Bai, Y.、Narayanasamy, P.、Garg, N. 和 Roy, N. SING：大语言模型中的空间上下文与下一代可穿戴设备。在 第四十二届国际机器学习会议。

Norval, M. 和 Wang, Z. 语音情感识别的可解释人工智能技术：聚焦于可解释人工智能模型。《style id='1'>《人工智能》，28(76)：85–123，2025年。

Pescatore, C. R., Zhang, H., Hadjinicolaou, A. E., Paulk, A. C., Rolston, J. D., Richardson, R. M., Williams, Z. M., Cai, J. 和 Cash, S. S. 利用人类颅内脑电图从自然语音中解码语义。*bioRxiv*，2025年。

Plantinga, P., Chen, J.-K., Sattari, R., Ravanelli, M. 和 Klein, D. 从黑箱到生物标志物：用于解读帕金森病语音模型的稀疏自编码器。*arXivpreprint arXiv:2507.16836*，2025年。

Saeidi, M.、Karwowski, W.、Farahani, F. V.、Fiok, K.、Taiar, R.、Hancock, P. A. 和 Al-Juaid, A. 使用机器学习对脑电信号进行神经解码：系统综述。《脑科学》，11(11):1525，2021年。

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Or- tega, A., Bloom, J. 等。机制可解释性中的开放问题。*arXiv*预印本 *arXiv:2501.16496*，2025年。Shimizu, R., Antonello, R. J., Singh, C. 和 Mesgarani, N.。语音的可解释嵌入可增强并解释音频模型的脑编码性能。*arXiv*预印本 *arXiv:2507.16080*，2025a。Shi mizu, R., Antonello, R. J., Singh, C. 和 Mesgarani, N.。语音的可解释嵌入可增强并解释音频模型的脑编码性能。*arXiv*预印本 *arXiv:2507.16080*，2025b。Sujatha Ravindran, A. 和 Contreras-Vidal, J.。使用模拟真实值对脑电图深度学习可解释性方法进行的实证比较。《科学报告》，13(1):17709，2023年。Williams, I., Oldenburg, N., Dhar, R., Hatherley, J., Fierro, C., Rajcic, N., Schiller, S. R., Stamatiou, F. 和 Søgaard, A.。机制可解释性需要哲学。*arXiv*预印本 *arXiv:2506.18852*，2025年。Wu, X., Wellington, S., Fu, Z. 和 Zhang, D.。使用先进的深度学习方法从立体脑电图（seeg）信号进行语音解码。《神经工程学杂志》，21(3):036055，2024年。Zhang, F. 和 Nanda, N.。迈向语言模型中激活修补的最佳实践：指标与方法。*arXiv*预印本 *arXiv:2309.16042*，2023年。Zheng, H., Wang, H.-T., Jing, Y.-T., Lin, P.-Y., Zhao, H.- Q., Chen, W., Wei, P.-H., Shan, Y.-Z., Zhao, G.-G. 和 Liu, Y.-Z.。Brainstratify：颅内神经动态的由粗到细解耦。*arXiv*预印本 *arXiv:2505.20480*，2025年。

A. 神经语音　　发声、想象和默念语音模式下的解码器性能　　s

在各模式中，基线解码器表现出明显的性能差异，且在各项评估指标上表现一致（图6）。发声语音实现了最强的重建质量，具有最高的皮尔逊相关系数和动态时间规整皮尔逊相关系数，以及最低的频谱失真（梅尔倒谱失真）和基频误差（基频均方根误差）。想象语音相对于发声语音有所下降，但在某些基于相关性的度量上仍具可比性，并且在样本间表现出更高的变异性。模拟语音总体上始终是最具挑战性的条件，其皮尔逊相关系数/动态时间规整皮尔逊相关系数最低，梅尔倒谱失真和基频均方根误差最高。这些基线促使我们进行因果干预，以检验跨模式差异是否可归因于内部表征而非输入差异。

表 2 报告了不同语音模式下的基线重建质量（均值 ± 标准差）。发声模式的整体性能最佳，想象模式居中但方差较大，而默念模式始终是最具挑战性的条件。

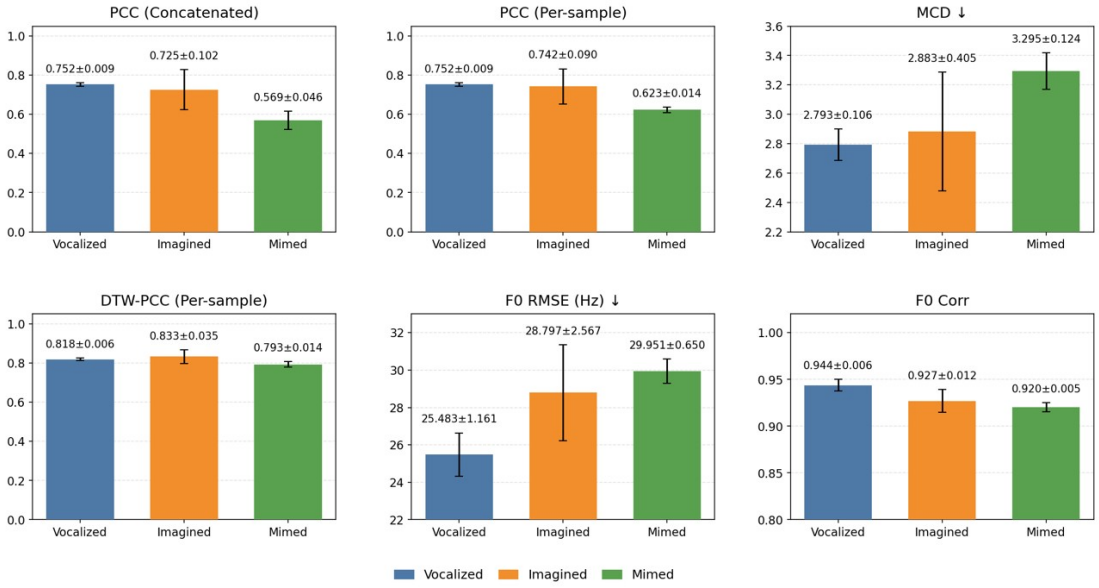


图*6.* 不同语音模式下的基线解码性能（均值 ± 标准差）。PCC/动态时间规整皮尔逊相关系数/基频相关系数越高越好；梅尔倒谱失真/基频均方根误差越低越好。

指标	发声	想象	默念
皮尔逊相关系数（拼接）↑	0.7519 ± 0.0089	0.7254 ± 0.1021	0.5686 ± 0.0459
皮尔逊相关系数（逐样本）↑	0.7519 ± 0.0089	0.7419 ± 0.0903	0.6231 ± 0.0144
梅尔倒谱失真 ↓	2.7928 ± 0.1056	2.8830 ± 0.4053	3.2948 ± 0.1236
动态时间规整皮尔逊相关系数（逐样本）↑	0.8184 ± 0.0064	0.8326 ± 0.0352	0.7928 ± 0.0143
基频均方根误差（Hz）↓	25.4827 ± 1.1614	28.7965 ± 2.5670	29.9511 ± 0.6502
基频相关系数↑	0.9438 ± 0.0063	0.9271 ± 0.0120	0.9204 ± 0.0048

表*2*。 不同语音模式下的基线性能（均值 ± 标准差）。皮尔逊相关系数/动态时间规整皮尔逊相关系数/基频相关系数越高越好，梅尔倒谱失真/基频均方根误差越低越好。

B. 作为神经语音脑机接口基础的机制理解

B.1. 基线解码性能与全激活修补

在基线设置下，不同语音模式的解码性能呈现出清晰的层级：发声语音解码准确（皮尔逊相关系数 $= 0.752$，梅尔倒谱失真 $= 2.8$），想象语音的表现居中（皮尔逊相关系数 $= 0.725$，梅尔倒谱失真 $= 2.88$），而默念语音的解码效果较差（皮尔逊相关系数 $= 0.5686$，梅尔倒谱失真 $= 3.34$）。

全激活修补表明，这一层级完全由模型内部表征决定。将想象或默语输入的内部激活替换为发声语音的内部激活后，发声水平的

条件	皮尔逊相关系数 ↑	梅尔倒谱失真 ↓
基线解码		
发声	0.752	2.8
默念	0.5686	3.34
想象	0.725	2.88
全激活修补		
发声 → 想象	**0.954**	**1.63**
发声 → 默念	**0.954**	**1.63**
默念 →想象	**0.534**	**3.31**
想象 → 发声	**0.177**	**3.70**

表*3.* 基线解码性能和全激活修补结果。完整补丁将输入模式的卷积或循环激活替换为供体模式的相应激活。

表现得以恢复并提升。反之，强制发声输入使用想象语音激活，会使表现降至想象水平，并进一步降低皮尔逊相关系数。这种精确的数值对称性同时证明了因果充分性和必要性：高质量解码依赖于维持发声内部表征，而无论输入的脑电信号如何，替换这些表征都足以引发或破坏解码准确率。

表 3所示的这些二元修补结果促使我们进一步开展更细粒度的分析，以探究因果信息如何在内部表征中分布和插值。

C. 三模态激活插值

在所有三模态配对中，激活插值使解码质量平滑且单调地变化，并且在不同插值方向上呈现对称行为。这表明，想象、默念和发声语音并非以离散类别的形式表征，而是沿着模型内部共享的连续因果轴分布（图 7）。

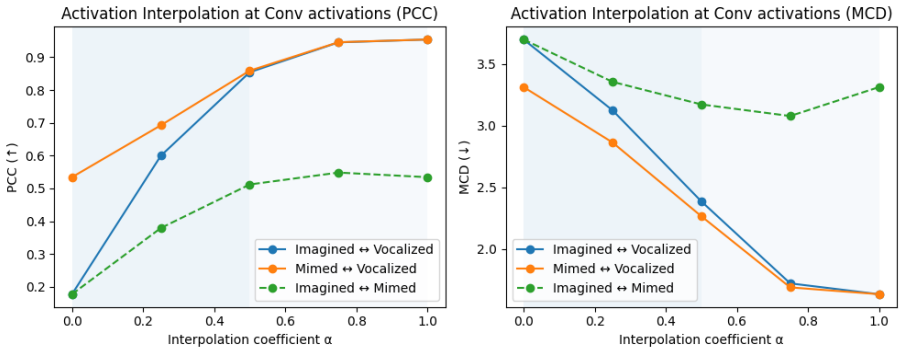
插值效应因层而异。在卷积层，皮尔逊相关系数和梅尔倒谱失真均随插值系数近乎线性地变化，这表明局部发音和频谱结构以平滑的组合方式编码。相比之下，循环层的插值表现出较不稳定的中间状态，这说明循环表征编码了更高层次的时间组织，需要更强的对齐才能实现稳定的解码性能提升。

D. 因果定位与子空间集中

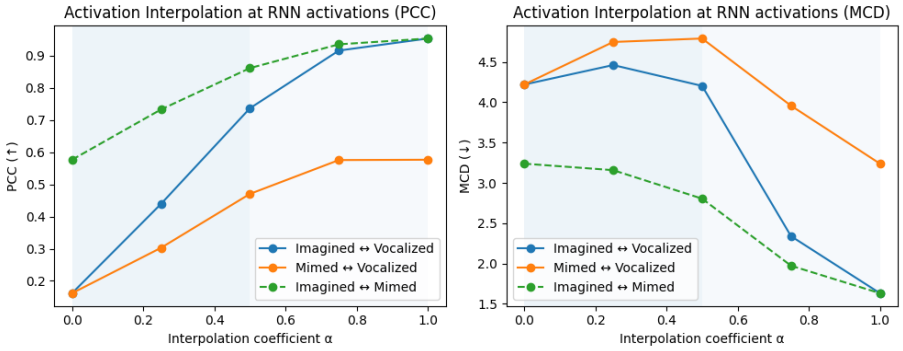
定向激活修补将跨模态因果信息定位到特定的内部区域，如表 4 所示。粗略分组和滑动窗口干预揭示，循环状态中存在一个狭窄的时间瓶颈，且少数卷积特征通道主导了迁移效应。综合来看，这些结果表明，跨模态语音解码依赖于结构化的低维内部子空间，而非均匀分布的表征。

条件	卷积 g0	卷积 g1	卷积 g2	卷积 g3	RNN早期	RNN中期	RNN晚期
充分性：想象 ← 发声							
PCC	0.576	0.535	**0.659**	0.503	**0.450**	0.427	0.402
MCD	3.23	3.16	**3.00**	3.21	3.29	3.37	3.41
必要性：发声 ←想象							
PCC	0.940	0.935	0.932	0.934	**0.698**	0.694	0.686
MCD	1.76	1.79	1.82	1.83	**2.30**	2.41	2.82

表*4.* 由粗到细的因果定位结果横向重新排版。我们报告选择性修补卷积通道组或RNN时间窗口时的解码性能。充分性对应于想象 ← 发声修补；必要性对应于发声 ← 想象修补。



(a) 卷积层



(b) 循环层

图7. 跨模型层的三模态激活插值。解码性能随插值系数 α 平滑变化，模拟语音始终处于中间表征位置。

E. 神经元级饱和分析中的模态依赖性不对称

性能变化的幅度和方向也很大程度上取决于源模态。当源模态为发声时，即将发声模型的激活修补到默念或想象目标中（图 5 中的绿色曲线），我们观察到皮尔逊相关系数有最大幅度的提升。这一效应与发声脑反应和语音的声学结构相关性更强有关，因为发声会同时激活运动活动和听觉反馈。相比之下，当目标模态为发声时，从默念或想象激活进行修补几乎不会提升皮尔逊相关系数，甚至毫无提升（图 5 左侧最边缘的面板），这表明解码器对发声语音的处理已接近最优状态。

在考察模拟与想象模态之间的交互作用时，还发现了另一种有趣的模式。想象 →默念干预（图5，中间面板）显示，皮尔逊相关系数（PCC）先上升，随后下降，在卷积层中尤为明显。这一提升反映了默念与想象之间共享的神经过程，可能与发音意图或语音规划等内部机制有关。相比之下，默念 → 想象干预（图5，最右侧面板）带来的 PCC 提升很小，表明这两种模态之间的跨模态激活传递存在不对称性。

F. 神经元补丁的句子级分析

图 8 展示了跨模态激活修补的句子级获胜者覆盖曲线。在所有目标（被修补至）模态中，覆盖率随 k增加，这表明不同句子依赖于不同的神经元，而非单一的小型神经元集合。

覆盖率曲线的上升速度揭示了因果贡献是局部化还是分布式。图 8 中，上升越快，表明行为越局部化，即少数神经元便能在许多句子中带来 PCC 提升；而上升越慢、越平缓，则表明神经元参与更加分布式。

首先观察到，当目标模态为默念时，覆盖率曲线的上升最为陡峭。这表明，当目标为默念时，关键神经元在不同句子间更加一致，尤其是在从发声模态进行修补时。这一模式与图 5中的饱和曲线一致：从发声模态进行修补时，默念目标的 PCC 提升最大。从神经科学的角度来看，模拟语音与发声语音具有相同的运动结构，而解码器似乎会重复利用共同的发音子空间，这在此表现为更高的句子级覆盖率。

另一项观察与模型架构有关。卷积层中的神经元（图8中的虚线）在 k 的所有取值上，覆盖率始终低于循环层中的神经元。这表明关键神经元在卷积层中分布更为分散，而循环层则依赖一组较小的、可跨句子泛化的神经元。这说明循环层编码了更多句子级抽象，而卷积层捕捉的是更局部、更具句子特异性的信息。

最后，覆盖率曲线的形状取决于源-目标模态对。当源模态为发声时（图8中的绿色曲线），覆盖率上升得更快。尤其是在发声 →默念条件下，曲线上升最为陡峭，前五神经元覆盖了约35%的句子集合。这表明发声表征在迁移到其他模态时会激活更为一致的神经元子集。相比之下，当目标模态为发声时，覆盖率增长更慢且几乎呈线性，说明PCC提升来自一组更大、分布更广的神经元。

表5汇报了所有源-目标模态对和层中句子级获胜一致性的定量汇总。

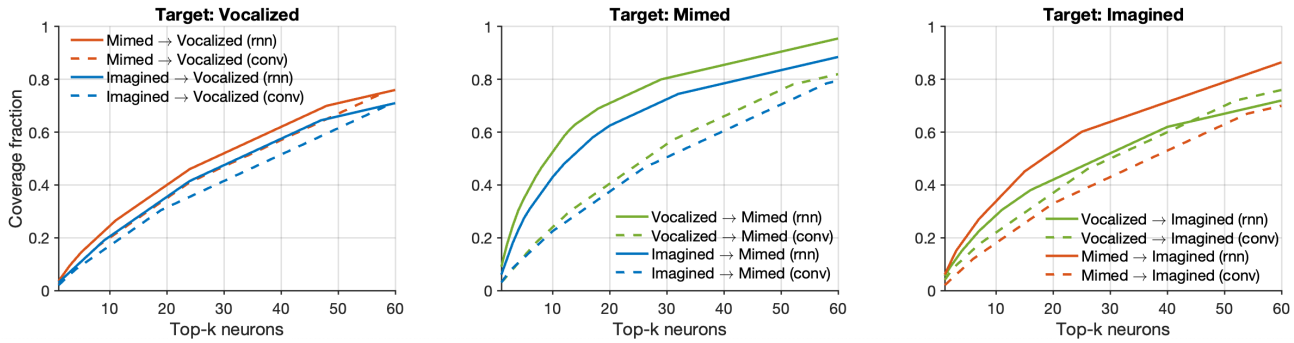


图*8.* 跨模态激活修补的句子级获胜者覆盖曲线展示了三种目标模态：发声（左）、默念（中）和想象（右）。每条曲线表示从源模态进行修补后，正确目标在前 k 个神经元中排名的句子所占比例。

源 → 目标	层	句子数	唯一数	前1（%）	前5（%）	熵
想象 → 默念	conv	200	101	3.00	12.50	6.45
	rnn	200	83	6.00	27.50	5.77
想象 → 发声	conv	200	118	2.00	9.50	6.73
	rnn	200	118	3.00	11.00	6.66
默念 → 想象	conv	200	120	2.00	10.00	6.73
	rnn	200	129	4.00	14.50	6.73
默念 → 发声	conv	200	108	2.50	11.00	6.57
	rnn	200	108	3.50	14.50	6.49
发声 → 想象	conv	200	108	4.00	13.50	6.52
	rnn	200	116	5.50	17.50	6.51
发声 → 默念	conv	200	96	3.00	13.00	6.34
	rnn	200	69	**9.00**	**35.00**	**5.35**

表 5. 单神经元激活修补的句子级获胜者一致性统计。我们报告唯一胜者神经元的数量、前一和前五覆盖率（由出现最频繁的获胜神经元所覆盖的句子比例），以及获胜者分布的熵。所有结果均基于每个条件的 200 个句子计算。

G. 为什么可解释性对面向神经科学的模型至关重要

深度神经网络的设计受到大脑工作方式的启发，其构建目的是以某种类似于大脑处理过程的方式应对复杂任务。尽管这些模型并非旨在精确复制生物机制，但它们有助于检验某些计算结构或模式是否与我们在神经数据中观察到的相符。

可解释性方法，尤其是机制可解释性，有助于我们超越模型的性能，了解训练后的神经网络与大脑有多相似。这些方法不再将模型视为黑箱，而是考察其内部模式如何组织、信息如何在其中传递，以及哪些部分负责重要行为。在神经科学领域，这使我们能够将模型的内部运作与神经系统已知的特征进行比较，例如分布式模式、瓶颈和组织层级。

从这个角度看，可解释性实验就像是在向模型提出问题。它们让我们得以 выяснить，网络解决任务时采用的是类似大脑的策略，还是使用了虽能得出正确答案、却有所不同的捷径。这一区别很重要，尤其是在模型使用神经数据进行训练时，因为仅有高准确率并不能保证模型的模式准确反映大脑的工作方式。

这项工作也有实际意义。生物系统具有很强的鲁棒性，能够适应不同的情境和噪声水平。如果某些特征有助于提升这种鲁棒性，那么组织结构更接近大脑的模型也可能适应得更好，尤其是在数据较少或条件不断变化时。可解释性技术通过将模型的内部结构与其泛化能力联系起来，帮助检验这一观点，而不只是考察模型的输入和输出。

在神经科学中，它为研究人员提供了工具，以检验关于他们能够控制和改变的系统中计算过程的设想。在机器学习中，可解释性有助于识别哪些内部结构有助于对神经数据进行建模，从而让神经数据直接影响模型的构建方式。