《心理科学进展》审稿意见与作者回应

题目：视听言语感知的贝叶斯因果推断机制
作者：董晨杰，王穗苹

第一轮

审稿人意见：

本文关注视听言语感知这一人类重要的信息处理能力，拟借助贝叶斯因果推理框架，考察视听言语感知中不同抽象水平信息的整合机制。研究问题较明确，研究思路清晰。关于文章，我有如下几个问题想与作者讨论：

意见 1：前言中说“前人研究已就特定脑区定位达成共识”，但研究一的目的仍是确定视听感知脑区。这两者是否矛盾？前人研究是否有关键性的问题尚未解决？

回应：谢谢审稿人对论文的积极评价和宝贵建议。前言所述内容与研究一属互补关系，具体如下。

前言提及“前人研究已就特定脑区定位达成共识”主要指对于“颞上沟后侧（pSTS）的多感觉整合脑区”的定位。然而，对于（1）pSTG/S 涉及的哪些功能亚区？除了 pSTG/S 外，视听言语信号的处理还涉及哪些脑区；（2）以 pSTS 为核心的大脑视听言语加工网络，采用何种计算机制处理视觉、听觉信号？这两个问题仍尚未解决。为解答这两个问题，研究一基于“经典视听操纵结合条件间比较”、“生态化的全脑编解码”两种研究视角和分析思路，更系统地界定与视听言语感知相关的多感觉模态整合脑区。

为提高论文内容的连贯性和可读性，减轻读者疑惑。我们在修改稿中（第 7 页：第 176 行-第 179 行；第 8 页：第 203 行-第 204 行；红色标注），增加了更明确表述。

- **第 176 行-第 179 行：***“针对第一个问题，研究一将系统界定视听言语感知的大脑网络（图 2）。具体而言，我们将在前人已界定的 pSTG/S 多感觉整合区的基础上 (Deen et al., 2023; Pitcher & Ungerleider, 2021)，揭示该脑区包含的功能亚区，界定其他参与视听言语整合的脑区，考察这些脑区的具体功能与信息编码模式。”*

- **第 203 行-第 204 行：**“*研究一旨在界定负责视听言语感知的大脑网络，深入分析包括 pSTG/S 在内的视听整合脑区的具体功能操作与信息编码模式，包括两个 fMRI 实验。*”

意见 2：“抽象水平”具体是怎么定义的？看前言和研究一似乎是不同类型的信息（“例如物理层面的时空线索、社交相关的讲话者身份特征、语言特异的音节和语义特征等”），但研究二中似乎强调抽象程度的高低。建议有一个明确的定义。

回应：谢谢您的建议！确实，在初稿中，我们未明确定义“抽象水平”这一概念，并在部分段落与“抽象程度”交替使用。在本文中，“抽象水平”是指言语感知的过程中，各类信息所包含的抽象语义概念的程度。例如，时间（或空间）信息涉及的语义概念低于讲话者身份、音节等信息。由于这些信息可通过多个感觉模态被处理，因此，在研究二中，我们提出视听关系也具有不同的“抽象水平”，个体可基于时空、身份、语义等（抽象程度从低到高的）视听关系，推断视听信号的来源，并基于此整合或分离信息。

针对该建议，作者已在修改稿中（第 3 页：第 73 行-第 79 行；红色标注），增加了对“抽象水平”的定义，并在全文统一使用该表述。

- **第 73 行-第 79 行：**“*言语信号可传递多类信息(Holler & Levinson, 2019)，这些信息包含的抽象语义概念的程度（下文简称为“抽象水平”）存在差异。例如，个体可从视听言语信号中获得物理层面的时空线索、社交相关的讲话者身份信息、语言特异的音节和语义信息等，它们的抽象水平逐渐增加(Mesgarani et al., 2014; Bhaya-Grossman & Chang, 2022; Gao et al., 2024)。更为重要的是，上述信息均可被听觉和视觉单独处理，也可通过多个感觉模态进行加工，形成抽象程度各异的、复杂视听对应关系(Spence, 2011; Young et al., 2020)，影响对言语的感知。*”

意见 3：研究一的关键变量“多维度特征”具体包含哪些特征？为什么这些特征是重要的。

回应：谢谢，这个问题对于单体素编码研究设计、数据分析非常重要。在我们的设计中，研究一至少纳入三个维度的特征，具体如下。（1）视听刺激包含的时间、空间，以及运动等物理特征，例如音频包络的变化、画面中的朝向变化等。（2）视听刺激传递的多模态社交特征，例如音频、画面中人物身份、情绪类别特征。（3）视听刺激涉及的抽象语义特征，例如音频、画面中提及的特定概念等。

这些特征的重要性体现在两个方面，（1）它们均具有多感觉模态特性，在听觉模态和视觉模态表现出对应性。（2）这些特征与视听言语感知中信息的抽象水平相对应，个体可以从特定维度或结合多个维度的特征，推断视听信号的来源，影响视听言语感知的结果。

我们已在修改稿中（第 9 页：第 220 行-第 225 行；红色标注），增加了对应的表述。

- **第 220 行-第 225 行：**“*我们将至少纳入三个维度的特征：（1）视听刺激包含的时间、空间，以及运动等物理特征，例如音频包络的变化、画面中的朝向变化等。（2）视听刺激传递的多模态社交特征，例如音频中的音色身份、情绪类别特征、画面中的面部身份、情绪类别特征。（3）视听刺激涉及的抽象语义特征，例如音频、画面中提及的特定概念等。这些特征均具有多感觉模态特性，并与视听言语感知中信息的抽象水平相对应，个体可以从特定维度或结合多个维度的特征，推断视听信号的来源，影响视听言语感知的结果。*”

意见 4：文章里没有图 3.

回应：抱歉，此处排序错误，实为图 2。修改稿已更正所有图表序号和引用。

意见 5：此外，建议前言中以表格或单独一段的形式整理研究问题、研究假设和预期，便于读者理解

回应：感谢，该建议对于提高论文可读性极有帮助！我们已于修改稿的“2 研究现状评述”的小结部分，即在介绍完项目相关的研究背景后（第 6 页：第 156 行-第 166 行；红色标注），补充相关内容。

- **第 156 行-第 166 行：**“*具体而言，本方案将系统界定视听言语感知的大脑网络，深入分析该网络中核心脑区所参与的计算过程。研究一将界定 pSTG/S 多感觉脑区的功能亚区，揭示其他参与视听言语整合的脑区，并考察这些脑区的具体功能。我们预期 pSTG/S 的功能亚区分别参与对不同的视听关系加工，例如，视听同步性涉及颞上回/沟前部、视听身份对应性涉及颞上回/沟后部。视听言语信号中不同抽象水平的视听信息，分布式编码于感觉模态特异脑区、多感觉脑区，以及超模态语言网络。研究二基于 BCI 框架，考察不同抽象水平的视听关系如何影响言语感知，及其神经基础。我们预期，具有较高感觉模态特异性的脑区将根据 FS 模型独立表征视觉、听觉信息；位于 pSTG/S 的视听语音加工区将遵循 FF 模型，整合视听语音信息；基于视*

听关系的因果推断过程，独立地表征于具有不同功能选择性的多模态区域；IFG 基于视听关系的抽象水平，调节不同因果推断在视听言语感知中的交互过程。”

第二轮

审稿人意见：

我认为作者已经很好地回应了我先前的问题和疑惑：1. 两项研究的创新点和关键问题均已突出；2. 关键性的概念和操作变量均已清晰定义；3. 通过对拟考察问题及其预期的归纳整理，文章的可读性明显提高。

编委 1 意见：已审阅全文，同意发表