《心理科学进展》审稿意见与作者回应

题目：嗓音模仿认知神经加工的多阶段模型
作者：胡砚冰，蒋晓鸣

第一轮

审稿人 1

意见 1：这篇综述主要从三个方面探讨了嗓音模仿的认知加工机制。三个部分大概可总结为认知模型、神经基础和个体差异表现。嗓音模仿的发生涉及一系列复杂的感觉运动整合过程，作者仔细地划分并探讨了这些整合过程，这对进一步分析并理解嗓音模仿行为十分重要。
回应：感谢审稿人对本综述所探讨问题重要性的认可。我们仔细地针对您提出的每一条建议进行了修改和回应，以进一步提高文章质量。为方便您的查阅，所有对应的修改和回应已在正文中用蓝色标注。

意见 2：在探讨认知模型部分，作者提及了言语感知-产出整合模型和关联序列学习模型。其中，言语感知-产出整合模型认为嗓音模仿的发生是因为说活人在言语感知过程中，对目标说话人的讲话内容进了预测；该预测是基于说话人自身的言语产出系统完成的，特别是前向模型的参与。作者提出了整合模型是从嗓音声学特征角度来阐明嗓音模仿的认知过程，这一点很重要，但前向模型具体是如何帮助完成语音信息的分析和计算从而引导发声动作的改变也许可以进一步阐明。前向模型在言语感知-产出整合模型中是很重要的一部分，特别是对于声学特征的分析和计算。
回应：感谢审稿人的建议。

首先，我们在文中明确了，说话人基于自己的产出指令来完成预期目标说话人下一步所说的内容，背后的重要认知过程是基于前向模型的参与。我们添加了相关内容（详见 p1；引言第 2 段)：“前向模型(forward model)在这种引导过程中起着核心作用(Pickering & Garrod, 2013)。前向模型可以理解为说话人准备说话之前就预期了其嘴唇、舌头和其他发声器官应该如何移动，以便产生期望的嗓音效果(蔡笑, 张清芳, 2020)。”

其次，我们补充了相关内容，分别来澄清前向模型和逆向模型在嗓音模仿过程中的计算过程（p5；第 2 段)：“具体来说，在嗓音模仿过程中，前向模型负责预测发声器官（如嘴唇和舌头）应如何运动以生成预期嗓音的声学特性。一旦嗓音生成，基于预期的发声动作及其后果与实际输出会进行比对，使前向模型能实时调整发声器官的动作以更精准地接近目标嗓音。相对于这一过程，逆向模型则用于发声动作的复制。它根据目标说话人的发声动作和声学特征来生成相应的发声器官运动参数，从而使说话人能够执行与目标说话人相似的发声动作。这两个模型共同协作，确保了声学特性的再现和发声动作的精确复制。”

意见 3：另外，关联序列学习模型则强调了感觉-动作联结（sensorimotor link）的建立，侧重于逆向模型在模仿过程中的作用，这在第二部分（神经机制）部分有着重讨论。但有一点是，关联序列学习模型强调的是感觉和动作事件之间的联系（因为两个事件经常在时间上临近发生，即联结经验），而联结的两个事件之间是否一致或具有模仿性（matching or imitative）并不重要。这在嗓音模仿过程中，特别是当模仿者要对不熟悉的发音进行模仿

时，已有的感觉-动作联结可能反而会成为阻碍。这里就涉及到了后文提及的模仿行为的自发性和目标性。

回应：感谢审稿人的建议。我们认同您提到的关于感觉-动作联结在模仿不熟悉听觉事件时可能成为阻碍的观点。然而，如果这种新的听觉事件和已有的联结经验是相似的，这可能促进模仿者去模仿这一听觉事件。综合这些内容，我们加入了 ASL 来解释刺激-反应一致性结果的相关内容，提出了这种联结经验带来的行为后果会受到新感觉事件熟悉性的调节。我们添加的内容详见 p5（第 1 段）：“基于 ASL 模型，我们可以对刺激-反应一致性效应进行深入解释。ASL 模型特别强调感觉与动作之间的联结，这种联结会因为两者在时间维度上的邻近而得到加强。在这一框架下，如果一个新的感觉事件与已有的联结经验相似，那么它更可能促进相应的动作产生；相反，如果新的感觉事件与联结经验不吻合，那么它可能会妨碍动作的产出。举个实际例子，我们在模仿母语的听觉事件时通常会比模仿一种不熟悉的语言更为得心应手。”

意见 4：基于跟读范式和刺激反应一致性范式的结果，作者指出了嗓音模仿行为的自发性。一个行为的自发性可以是多维度的，其中一个维度是 unintentionality，即该行为的发生不受运动者本人意图的影响，特别是在刺激反应一致性范式里（see Ramsey et al., 2019）。作者引用了(Dufour & Nguyen, 2013)的研究来说明，不管模仿意图与否，行为上的聚合效应（即嗓音模仿）是无差异。此外，作者引用的(Garnier et al., 2013)fmri 结果也表明，模仿与跟读任务在相同程度上涉及了大脑背侧感觉-运动网络；但该实验的行为结果显示模仿与跟读任务的行为表现是不同的。在后文提及的 fmri 研究里，有不少研究采用了有意模仿任务，作者也提出了嗓音模仿行为的目标性，也许作者可以进一步解释一下自发性和目标性是如何兼容，在概念上会有一些冲突，特别是考虑到模仿行为自发性里包含了 unintentional 这一特点。也许两个性质兼容，且在不同情况下（或不同任务中）参与度不同，但可以一起引向模仿行为。对自发性和目标性的进一步阐明也许可以让认知模型部分与神经基础部分之间的讨论更紧密一些。

审稿人提及参考文献：

Ramsey, R., Darda, K. M., & Downing, P. E. (2019). Automatic imitation remains unaffected under cognitive load. Journal of Experimental Psychology: Human Perception and Performance, 45(5), 601–615. https://doi.org/10.1037/xhp0000632

回应：感谢审稿人的建议。我们认同您对于当前稿件中有关自发性的定义存在不准确的意见。根据您的推荐，我们阅读了 Ramsey et al.(2019)的研究，对于当前嗓音模仿中自发性的界定很有启发。具体来说，感觉-联结的自动化反应和无意图的声学再现可以认为是自发性加工的两种维度（Ramsey et al., 2019）。为此，我们认为自发性可以进一步细分为两个层面：一是基于感觉-动作联结的自动化反应（通常观察于刺激-反应一致性范式中），另一是即便在无明确意图的情况下仍能准确再现声学特性（主要基于跟读范式的研究结果）。我们在两个部分添加了相关内容。

第一个部分在 p5，第 2 段：“嗓音模仿所涉及的两种自发性也存在区别：在刺激-反应一致性范式中，自发性主要表现为基于感觉-动作联结的自动化反应。这意味着模仿行为几乎是一种由刺激触发的自动反应。与之不同，跟读范式中的自发性更侧重于无明确意图下的声学特性再现。也就是说，即使没有明确的模仿意图，说话人仍然能准确地再现目标说话人的声学特性。”

第二个部分在 p8，第 2 段：“如前文所述，嗓音模仿具有两个核心特性：自发性和目标性。自发性可以进一步细分为两个层面：一是基于感觉-动作联结的自动化反应（通常观察

于刺激-反应一致性范式中），尽管以往的成像研究并未直接针对使用刺激-反应一致性范式来研究嗓音模仿中自动化加工特性的相关神经机制，但手势动作模仿的成像证据仍可提供有用的参考，原因在于这些手势动作模仿的研究与刺激-反应一致性范式都是基于 ASL 理论模型的假设(Cracco et al., 2018)。这些研究发现观察和执行动作的过程会激活额下回和初级运动皮层，这些区域都与镜像神经元系统有关(Cracco et al., 2018)。镜像神经元系统，尤其是在猕猴脑中的前运动皮层的 F5 区域，被认为是模仿和语言发展的神经基础(Nguyen & Delvaux, 2015)。这一系统通过促进观察到的动作和声音的内部映射，为一般性的模仿行为提供了神经基础。在嗓音模仿中，这些镜像神经元就会启动。它们不仅帮助说话人准确地“听”到目标说话人嗓音的特点，还将这些信息转换为具体的发声指令，好让说话人的喉咙和嘴巴知道要怎么动才能模仿出相同的声音。另一是即便在无明确意图的情况下仍能准确再现声学特性（主要基于跟读范式的研究结果）。在目标性方面，模仿行为不仅是一种无意识的反应，也是一个有目标的过程。在模仿过程中，说话人通过逆向模型生成实现预定目标状态所需的动作指令。根据当前关于嗓音模仿的神经机制研究，我们发现无论是有意图的模仿还是无意图的模仿，这两种不同类型的模仿涉及的脑区都是相关的。这表明有意图模仿和无意图模仿在神经层面上可能共享相似的处理路径或网络。相应的行为证据发现，与非模仿条件相比，两种模仿方式都能导致声学特性的聚合。这意味着声学特性的再现可以是无意图、自发产生的，也可以是有意图、目标导向的。这两者的主要差异可能体现在声学特性再现的程度上。例如在无意图的模仿中，说话人可能会在快速的言语交流环境中与目标说话人在声学特性上逐渐接近和靠拢，以促进更有效的沟通和合作。而在有意图的模仿中，说话人可能会更加精细地调整，以消除自己与目标说话人在声学特性上的差异，从而更接近“声学特征完全相同”的目标。”

意见 5：在关于个体差异的部分，作者将嗓音模仿行为划分为了三个主要阶段（听觉感知，说话人嗓音感知映射产出表征以及嗓音产出）来具体讨论。引用的相关研究都很说明问题。但考虑到这篇综述里每个部分之间的紧密性，也许可以分别联系到前文提及的认知模型和神经基础。

回应：感谢审稿人的建议。我们基于当前稿件中有关认知模型以及神经基础的内容，重写了个体差异的结尾部分。具体来说，从言语感知－产出整合模型、ASL 模型以及每个认知过程对应的神经基础来解释了嗓音感知、感知映射产出，以及嗓音产出这三个主要加工阶段在嗓音模仿认知加工中的重要性。相关修改内容见 p11（第 2 段）：“综上，嗓音感知，感知映射产出，以及嗓音产出这三个主要加工阶段确实会影响嗓音模仿的行为后果。首先，在感知阶段，言语感知－产出整合模型与 ASL 模型都强调说话人需要准确地提取目标说话人有用的声学信息和可见的发声器官信息（如口型）。这些信息会影响后续的产出指令或发声动作。更准确地提取与目标嗓音相关的信息将有助于嗓音模仿的准确性。此外，这一过程与次级听觉皮层（如颞上回）有密切的关联。其次，在感知映射产出阶段，言语感知－产出整合模型指出，唱音障碍可能源于前向模型在预测发声器官的动作指令时存在的偏差。从 ASL 模型的角度来看，这种障碍也可能受到先前感觉-动作联结经验的影响。具体地说，不熟悉的听觉事件可能会受到已有感觉-动作联结经验的阻碍，而熟悉的听觉事件则可能受到这种经验的促进。这一认知过程与弓形束和基底神经节的活动有关。最后，言语感知－产出整合模型与 ASL 模型都强调了发声器官运动的灵活性。前者认为，为了满足前向模型预测的发声动作指令，需要足够灵活的发声器官运动。后者则指出，只有具备这种灵活性，才能丰富和扩展已有的感觉-动作联结经验。这一过程与初级运动皮层（如喉部，唇部以及舌部）的活动密切相关。通过这三个阶段的综合分析，不仅可以更全面地理解嗓音模仿的复杂性，还可以清晰地看到各种认知模型和神经基础如何共同作用于嗓音模仿。”

……

审稿人 2

意见 1：论文综述了嗓音模仿的认知神经机制，文章的重点是阐述嗓音模仿的认知神经多阶段加工模型，阐明嗓音模仿与感知-发生运动加工在神经机制上的联系和区别。论文所考察的问题具有一定的理论意义和实践价值。有以下建议供作者参考。

回应：非常感谢审稿人提出的宝贵意见。为了提升文章的质量，我们已经按照您的各项建议进行了认真的修改和补充。所有相关的改动和进一步的阐述在文章正文里都用蓝色字体进行了标注，以便于您的查阅。

意见 2：综述整体未反映出研究者对嗓音模仿加工机制的争论焦点，未突出强调嗓音模仿研究的重要理论意义。建议作者思考嗓音模仿研究的所属领域，将嗓音模仿的研究放在一个恰当的研究领域中进行介绍，在此背景下阐述嗓音模仿的研究能够解决的重要理论问题。

回应：感谢审稿人提出的宝贵意见。由于嗓音模仿涉及到言语感知与言语运动控制相关的认知过程，但是，相比于单纯的言语感知或言语产出过程，嗓音模仿需要考虑言语感知后果和言语产出信息之间的相似性，且具有自己独特的加工特点（如自发性和目标性)。为此，我们将嗓音模仿放在了言语交流中的言语感知运动控制（speech sensorimotor control）这一研究领域，并从言语感知运动控制所需要调用的认知过程出发，对嗓音模仿的认知加工机制进行阐述。

重要的是，言语感知运动控制通常带有一定社会目的。嗓音模仿的参与可以促进对话双方通过言语交流来达到一定的社会目的。因此，对嗓音模仿机制的研究可以揭示言语感知运动控制过程如何协助说话人实现特定的社会沟通。具体来说，嗓音模仿通过有社会目的的言语感知运动控制，可以让说话人在声学和发声器官动作两个层面与目标说话人表现出相似性。关键在于，嗓音模仿之所以能达成这些社会目标，是因为它具有自发性的特点。在这个背景下，我们基于以往关于言语感知运动控制相关的两个认知模型对嗓音模仿的认知加工机制进行解释。我们已对原始稿件内容进行了修改（详见 p1，引言第 1 段)：“言语交流不仅依赖于遵循特定的音系和句法规则(Chomsky & Lightfoot, 2002)，还有其深层的社会应用，如促进合作和情感联结。然而，仅凭固定的语言规则是不足以实现这些社会目标的。原因在于每个说话人都有独特的表达方式，这些方式反映了他们的人格特质和文化背景(Kinzler, 2021)。有研究指出，在以社会目的为导向的言语交流中，言语感知运动控制[1]机制(speech-sensory-motor control mechanism)起到了关键作用，在这一过程中，嗓音模仿的认知机制尤为重要，尤其是在推动对话双方在特定特征(如声学，语义，句法以及发声动作)上达到相似性时(Kinzler, 2021)。具体来说，随着对话的深入，双方会逐渐展现出在不同模态层面上的相似性，如声音和口型的同步。在这个模仿的过程中，说话人可能会借鉴他们感知到的目标说话人多模态信息，并控制自己的发声动作，以产出更接近目标说话人的嗓音，从而更有效地达到社会交流的目的(Bernhold & Giles, 2020; Heyes, 2021; Pardo et al., 2022; Pickering & Garrod, 2013)。嗓音模仿能有效地促进言语交流中的社会目的达成，其中一个关键因素便是其自发性的特点。这种自发性使说话人无需刻意模仿目标说话人的语言特征，而是通过内部机制（如前向和逆向模型）自然地将听觉信息映射为相应的发声指令。”在此基础上，我们对言语感知运动控制给了一个定义，并进一步明确了嗓音模仿于言语感知运动控制之间的区别和联系。相关内容，在 p1 (脚注 1)：“言语感知运动控制可以定义为一个包括听觉感知至发声（言语产出）的综合性认知加工过程(Bono et al., 2022）。这一过程是为了确保个体能准确地接收、理解以及回应言语信息。这一认知过程具体涉及听觉信息首先被感知和解析，

然后转化为一个产出运动计划，最终通过运动器官（如声带、舌头、嘴唇等）实现精准的言语产出。与言语感知运动控制不同，嗓音模仿要求说话人需要考虑如何让自己言语产出的信息与其感知到的目标说话人信息是相似的。”

在此背景下，我们明确了以往理论模型在解释嗓音模仿机制的不足以及需要解决的理论问题。我们已经补充相关内容(详见 p2，第 3 段)：“由此可见，从言语感知-产出整合模型的视角来看，嗓音模仿被视为说话人对目标说话人声学特性的再现。而从关联序列学习理论的角度看，嗓音模仿则更侧重于说话人对目标说话人发声器官动作的复制。两个模型在解释嗓音模仿方面都有其独特的优点和局限性。ASL 模型主要侧重于解释同一物种内基于发声器官动作的模仿机制，因此在处理跨物种模仿方面缺乏全面性。与之相反，言语感知-产出整合模型通过声学相似性来定义嗓音模仿，能够较好地解释跨物种的嗓音模仿现象，从而弥补了 ASL 模型在这方面的不足(Cracco et al., 2018; Mercado et al., 2014)。然而，ASL 模型提供了关于嗓音模仿形成机制的具体假设，特别是声学相似性是如何依赖特定发声器官来实现的，这是言语感知-产出整合模型尚未深入探讨的。总的来说，两个模型分别从嗓音中的声学信息和效应器官动作来探究说话人嗓音模仿的认知过程。然而，这种独立的观点忽略了一个事实：言语交流是一个涉及多模态信息输入和输出的复杂过程(Belyk, Eichert, et al., 2021; Belyk et al., 2019; Brown et al., 2021)。具体来说，嗓音模仿不仅依赖于说话人各种发声效应器官（如喉部、舌部、上颚部、唇部等）的协同作用，还需要通过这些器官产生的声学信号来模仿目标说话人。这表明嗓音模仿的认知过程不仅要求说话人精确地复制发声器官的动作表征，还需要再现与目标说话人近似的声音特征。现有的认知模型在两个方面都存在局限性：一是它们不能全面地解释嗓音模仿在多模态情境中是如何进行认知加工的；二是缺乏对嗓音模仿认知过程背后神经机制的明确解释。明确这些神经机制不仅有助于更深入地理解神经因素与模仿行为之间的因果关系，还可能为治疗发声障碍提供有临床意义的新视角。针对这些不足，需要当前研究梳理和整合以往的实证研究，以构建一个更为全面的嗓音模仿的认知神经加工模型。”

意见 3：第一段作者详细地描述了两个研究，其目的是什么？对于嗓音模仿机制的重要性体现在什么地方？在引言第一段如此详细地介绍两个研究是不合适的，建议作者在阐述这两个研究重要性的基础上进行有针对性的介绍。

回应：感谢审稿人提出的宝贵意见。在引言的第二段提到语义和句法模仿的目的，是为了更明确地突出嗓音模仿在认知加工方面的独特性。这是因为，与语义和句法不同，这两者往往是基于共同的社会或群体规则(如句法规则)，即便没有模仿，也自然会显示出一定程度的相似性。相反，嗓音的声学特性是个体特定且高度变异的，以至于人们可以仅通过嗓音来识别不同的个体。这高度的变异性使得嗓音模仿在模仿研究中的重要性尤为凸显，因为模仿加工能够使两个本来高度变异的嗓音变得更加相似。为了进一步突出嗓音模仿机制的独特性和重要性，我们修改了 p1-p2，引言部分的第 2 段：“与遵循语言规则的语义和句法不同，嗓音中的音段和超音段声学线索具有高度的灵活性和变异性，这使得个体间存在显著的差异。有研究进一步证实，说话人能够仅凭这些嗓音线索(如基频、共振峰等)轻松地识别不同目标说话人的身份(Perrachione et al., 2011)。模仿认知过程的介入可以有效地减少这些个体间的高度变异性，使得说话人在特定声学特征上更加接近目标说话人。因此，深入了解嗓音模仿的认知加工机制对于揭示言语交流中的社会合作行为具有重要的理论意义。”

我们有针对性的修改了两项研究。修改后的稿件，先介绍了与嗓音模仿有关的理论模型，并在有关理论模型下介绍具体的研究。具体修改内容在 p1，引言部分第 2 段：“言语感知–产出整合模型指出，说话人在语义、句法和嗓音这三个层面都会利用目标说话人的言语信息来调整自己的产出系统。具体来说，在语义层面，当说话人预见到将出现一个特定的单词（如

帽子：cap），他们会提前调整舌部位置以做好发音准备。如果实际出现的单词（如水龙头：tap）与预期不符，他们需要做出更多的舌部位置调整(Drake & Corley, 2015)。在句法层面，说话人会在他们的言语产出系统中预先设置预期的句法信息，比如冠词与名词的配对(Martin et al., 2018)。若这一产出系统受到任何形式的干扰，它将影响说话人对后续句法信息的正确提取和模仿。”

意见 4：言语整合产出模型和关联序列学习模型观点的差异需要有实验证据的支持，两类模型的主要争论是什么？为什么研究者要提出不同的模型？

回应：感谢审稿人的意见。目前据我们所知，没有一个很好的实证研究来直接表明两个观点的差异，为此本综述基于理论视角分别提出了两类模型的优势和缺点。两类模型的争论主要在于对嗓音模仿的操作性定义存在不同，基于言语感知-产出模型的观点主要界定嗓音模仿在声学维度上存在相似性，而关联序列学习模型主要界定嗓音模仿在感知和发声动作表征维度上的一致性。

我们补充了相关内容来明确两个模型的主要争论(详见 p2，第 3 段)：“由此可见，从言语感知-产出整合模型的视角来看，嗓音模仿被视为说话人对目标说话人声学特性的再现。而从关联序列学习理论的角度看，嗓音模仿则更侧重于说话人对目标说话人发声器官动作的复制。”

以往研究者提出不同的模型，其目的在于可以解释普遍的嗓音模仿现象，但其解释的方面各有其局限性。我们补充了相关内容(详见 p2，第 3 段)：“两个模型在解释嗓音模仿方面都有其独特的优点和局限性。ASL 模型主要侧重于解释同一物种内基于发声器官动作的模仿机制，因此在处理跨物种模仿方面缺乏全面性。与之相反，言语感知-产出整合模型通过声学相似性来定义嗓音模仿，能够较好地解释跨物种的嗓音模仿现象，从而弥补了 ASL 模型在这方面的不足(Cracco et al., 2018; Mercado et al., 2014)。然而，ASL 模型提供了关于嗓音模仿形成机制的具体假设，特别是声学相似性是如何依赖特定发声器官来实现的，这是言语感知-产出整合模型尚未深入探讨的。总的来说，两个模型分别从嗓音中的声学信息和效应器官动作来探究说话人嗓音模仿的认知过程。”

意见 5：作者强调：“当前综述基于言语交流背景，进一步从说话人角度梳理嗓音模仿的认知加工过程，并澄清这一过程如何受到目标说话人发声器官动作表征和嗓音声学特征的影响。”言语交流背景的特殊性体现在哪些方面？“说话人的角度”具体指什么？与听话人的角度相比其区别是什么？

回应：感谢审稿人的意见。在新的稿件中，我们对“言语交流背景”进行了明确，将嗓音模仿的认知加工过程置于“言语感知运动控制”的背景下进行探讨。在这一背景下，传统的语言理解过程并非当前综述关注的认知过程。相反，由于嗓音模仿机制的加入，说话人会在多大程度上受到目标说话人非言语线索（如发声行为伴随的动作）的影响。为了实现特定的社会交流目的，说话人需借助模仿的认知过程，以在某些特定特征上与目标说话人达到相似性。这样，我们的讨论不仅涵盖了言语交流的非言语方面，也突出了模仿认知在社会性言语交流中的核心作用。

我们已经修改了 p1，引言第 1 段：“言语交流不仅依赖于遵循特定的音系和句法规则(Chomsky & Lightfoot, 2002)，还有其深层的社会应用，如促进合作和情感联结。然而，仅凭固定的语言规则是不足以实现这些社会目标的。原因在于每个说话人都有独特的表达方式，这些方式反映了他们的人格特质和文化背景(Kinzler, 2021)。有研究指出，在以社会目的为导向的言语交流中，言语感知运动控制[1]机制(speech sensorimotor control mechanism)起到了关键作用，在这一过程中，嗓音模仿的认知机制尤为重要，尤其是在推动对话双方在特定特征

(如声学，语义，句法以及发声动作)上达到相似性时(Kinzler, 2021)。具体来说，随着对话的深入，双方会逐渐展现出在不同模态层面上的相似性，如声音和口型的同步。在这个模仿的过程中，说话人可能会借鉴他们感知到的目标说话人多模态信息，并控制自己的发声动作，以产出更接近目标说话人的嗓音，从而更有效地达到社会交流的目的(Bernhold & Giles, 2020; Heyes, 2021; Pardo et al., 2022; Pickering & Garrod, 2013)。嗓音模仿能有效地促进言语交流中的社会目的达成，其中一个关键因素便是其自发性的特点。这种自发性使说话人无需刻意模仿目标说话人的语言特征，而是通过内部机制（如前向和逆向模型）自然地将听觉信息映射为相应的发声指令。”

我们补充了对于“说话人角度”的解释，及其与“听话人角度”的实质区别。相关内容，我们已经补充到 p3，脚注 2：“说话人角度”具体是指在嗓音模仿的三个核心认知加工阶段(即嗓音感知、感知到产出的映射，以及嗓音产出)，都涉及到产出系统的参与。这与传统的“听话人角度”有明显区别。听话人通常更关注于是否成功地解码了接收到的信息，而说话人不仅解码信息，还进一步对这些信息进行深层次的编码(如说话人基于接受的信息，通过改变其发声运动行为，进而产出与接受信息相关的特定语言信息)，以实现特定的社会目的。两种主要的模型，即言语感知–产出整合模型和关联序列学习（ASL）模型，都强调说话人在感知阶段并不是被动的。相反，内部的产出系统在整个感知过程中起到了主动的作用。这意味着，即使在听或感知别人的言语信息时，说话人也在“内部地说”，尽管可能不会外显地产出。总体来说，从“说话人角度”出发研究嗓音模仿能够更全面地阐释其认知加工机制，特别是能更深入地理解嗓音模仿中的三个核心认知处理阶段。”

意见 6：在言语交流中，在倾听对方讲话时，其主要目的是理解别人的话语，之后根据理解产生对话交流。在此过程中一般不会进行重复跟读。综述中所提及的“在言语交流背景”下的嗓音模仿是什么意思？研究中所采取的任务与自然的言语交流是否存在不同？这种不同对实验结果和理论观点会产生哪些影响？

回应：感谢审稿人的意见。我们同意您的观点，即在常规的言语交流环境中，说话人通常不会外显地重复或模仿目标说话人的话。现有的综合研究表明，在听取对方发言时，说话人的内部产出系统实际上是参与言语理解的，即便这一过程并不需要外显地产出。这一机制在解释言语交流中如何实现流畅的话轮转换方面具有重要意义。我们对正文做了如下修改：

我们补充了内部产出系统参与言语交流相关内容(p3，脚注 2)：“此外，内部产出的机制在解释言语交流中如何实现流畅的话轮转换方面具有重要意义。相关证据进一步显示，话轮转换中的切换时间(大约 200ms，即在目标说话人刚结束发言后说话人开始发言的时间)要远小于一般图片命名任务中的反应时间(大约为 350ms)。这表明内部产出系统在控制言语交流节奏，特别是在话轮转换中，起到了关键作用。”

我们解释了“在言语交流背景”下的嗓音模仿，具体内容补充在 p1 第 1 段：“有研究指出，在以社会目的为导向的言语交流中，言语感知运动控制[1]机制(speech sensori-motor control mechanism)起到了关键作用，在这一过程中，嗓音模仿的认知机制尤为重要，尤其是在推动对话双方在特定特征(如声学、语义、句法以及发声动作)上达到相似性时(Kinzler, 2021)。具体来说，随着对话的深入，双方会逐渐展现出在不同模态层面上的相似性，如声音和口型的同步。在这个模仿的过程中，说话人可能会借鉴他们感知到的目标说话人多模态信息，并控制自己的发声动作，以产出更接近目标说话人的嗓音，从而更有效地达到社会交流的目的(Bernhold & Giles, 2020; Heyes, 2021; Pardo et al., 2022; Pickering & Garrod, 2013)。”

同时，本文也指出，与自然的言语交流不同，当前综述梳理的嗓音模仿实证研究还是具有不够生态的局限性，即目标说话人的嗓音信息都是经过实验室控制的，这与更加生态的自然言语交流还存在差别。这种不同可能会影响刺激-反应一致性的结果，因为这些研究都是

基于音节水平的证据来得出。这种不同也会影响跟读范式中的结果，原因在于跟读范式的任务是模仿相同的嗓音信息，而自然言语交流中目标说话人和说话人的嗓音信息可以“趋”同。然而，随着话轮的增加，说话人与目标说话人的嗓音信息可能会表现出增加的相似性，这值得未来研究进一步探究。以往关于嗓音模仿的认知模型对于解释自然言语交流存在局限性，其局限性表现在，自然的言语交流是多模态的。这一局限性也在当前理论模型中得到了解决。我们修改了展望的第 3 部分(p13，第 3 段)：“此外，嗓音模仿多阶段加工模型解释了在更自然或生态有效的言语交流中，嗓音模仿认知不仅仅是单一模态的，而是基于多模态信息进行的认知加工。这一观点有效地弥补了以往嗓音模仿实证研究中的一些局限性。例如，先前的研究在刺激-反应一致性范式中通常只关注音节水平，或者在跟读范式中，说话人跟读的内容与目标说话人完全一致。这些研究设计与真实世界中复杂、多模态的言语交流场景存在一定的距离。因此，未来的研究应当努力发展更具生态效度的嗓音模仿范式。具体来说，应在真实的言语交流环境中进行实验，并采用当前综述梳理的经典嗓音模仿测量指标，以更全面地了解说话人与目标说话人在嗓音模仿认知加工过程中的相互作用和影响。这不仅能提供更接近自然状态的认知模型，还有助于深化我们对嗓音模仿机制的理解。”

意见 7：欧氏距离差分数、声学聚合与嗓音模仿之间的关系需要更进一步详细的描述，声学聚合的认知含义是什么？

回应：感谢审稿人的意见。首先，我们在原稿件基础上对欧氏距离差分数、声学聚合与嗓音模仿之间的关系进行了更详细的阐释。其次，基于对以往研究的梳理，我们解释了声学聚合的认知含义，即它体现了一个深层次的信息处理机制，该机制使个体能够在听到某个目标声音后，无意识或有意识地调整自己的发音，以便与目标声音更为接近。因此，声学聚合可以视为社交认知功能的一个有用指标，它可能有助于促进社交互动，增强群体凝聚力，或者在更广泛的语境下，提高交流的效果和效率。

具体补充部分在 p4 第 2 段：“基于言语感知-产出整合模型对嗓音模仿的定义，涉及说话人可以再现目标说话人的声学特征。与这一定义密切相关的操纵性定义为声学聚合，即在对话或模仿过程中，一个人的声学特征（比如音高、音量或语速）逐渐变得更像另一个人。进一步来说，声学聚合揭示了说话人在社交互动或模仿活动中如何自然地调整自己的声音以适应或接近目标说话人。声学聚合可能反映出个体在社交互动中的适应性和合作倾向，用以促进社交凝聚或增强信息传递的效率(Pardo et al., 2022)。声学聚合的测量指标是欧氏距离差分数，即通过计算两个声音样本在多维声学空间中的“距离”来量化它们有多相似或不同。这个“距离”越小，说明两个嗓音样本越相似，也就意味着更强的声学聚合。Dufour 和 Nguyen (2013)的研究结果进一步说明了，在被要求模仿目标说话人和在自然跟读的情况下，声学聚合的程度是没有差异的。这可能意味着，不管是任务相关或无关地模仿，说话人都会在一定程度上模仿目标说话人的声音。”

意见 8：有关嗓音模仿的认知神经机制模型要更深入地阐述当前对这一问题的争论焦点，强调嗓音模仿认知神经机制研究的重要意义。

回应：感谢审稿人的意见。当前嗓音模仿的认知神经机制的争论焦点主要集中在两个方面，也是本综述试图澄清的。第一，需要澄清说话人是如何将多模态的嗓音感知与发声动作协同整合，以优化嗓音模仿的认知处理过程。第二，需要明确哪些特定的神经机制或神经网络在嗓音模仿不同阶段的认知加工过程中发挥着核心作用。

我们已经补充和修改嗓音模仿的认知神经机制模型对于解决这两个问题的贡献(p8，第 3 段)：“当前嗓音模仿的认知神经模型突出了多模态信息在嗓音模仿中的重要性，显示说话人需要通过整合嗓音中的听觉和视觉信息，以准确捕捉目标嗓音的独特特质。一旦这些信息

被准确捕捉，说话人还需要同步地调整自己的发声机制以达到最佳模仿效果。在这个复杂的认知加工过程中，STG–AF–IFG–M1 神经网络起到了中枢作用。这一网络专门负责从多模态输入中提取和解析关键的嗓音信息，并与发声机制进行有效的整合。此外，皮质下的纹状体（BG）和内侧膝状体区域进一步优化这一过程，它们协同输入和输出机制，参与动作序列的选择、协调和执行，以实现更精准和自然的嗓音模仿。这些组成要素相互作用，共同构成了一个高度复杂但协调的认知神经网络。”

其次，我们进一步强调了嗓音模仿认知神经机制研究的意义。相关内容补充在(p8，第 3 段)：“梳理嗓音模仿的认知神经机制为未来研究提供了两方面的启示。首先，这项研究有助于丰富了以往的认知模型，将具体的认知过程与特定的神经网络相对应。其次，通过明确嗓音模仿认知加工背后的神经基础，将为发声障碍患者和面临早期言语习得困难的幼儿提供了临床参考。”

意见 9：有关嗓音模仿的个体差异部分，根据仅有的几项研究得出的结论比较宽泛。

回应：感谢审稿人的意见。我们重新写了个体差异部分的结论部分。具体来说，我们基于前文所提及的认知模型和神经基础，来重新解释了这几项研究的结论，使其更加具体。

相关修改内容在 p11 的第 2 段：“综上，嗓音感知，感知映射产出，以及嗓音产出这三个主要加工阶段确实会影响嗓音模仿的行为后果。首先，在感知阶段，言语感知 - 产出整合模型与 ASL 模型都强调说话人需要准确地提取目标说话人有用的声学信息和可见的发声器官信息（如口型）。这些信息会影响后续的产出指令或发声动作。更准确地提取与目标嗓音相关的信息将有助于嗓音模仿的准确性。此外，这一过程与次级听觉皮层（如颞上回）有密切的关联。其次，在感知映射产出阶段，言语感知–产出整合模型指出，唱音障碍可能源于前向模型在预测发声器官的动作指令时存在的偏差。从 ASL 模型的角度来看，这种障碍也可能受到先前感觉-动作联结经验的影响。具体地说，不熟悉的听觉事件可能会受到已有感觉-动作联结经验的阻碍，而熟悉的听觉事件则可能受到这种经验的促进。这一认知过程与弓形束和基底神经节的活动有关。最后，言语感知–产出整合模型与 ASL 模型都强调了发声器官运动的灵活性。前者认为，为了满足前向模型预测的发声动作指令，需要足够灵活的发声器官运动。后者则指出，只有具备这种灵活性，才能丰富和扩展已有的感觉-动作联结经验。这一过程与初级运动皮层（如喉部，唇部以及舌部）的活动密切相关。通过这三个阶段的综合分析，不仅可以更全面地理解嗓音模仿的复杂性，还可以清晰地看到各种认知模型和神经基础如何共同作用于嗓音模仿。”

意见 10：总体感觉该篇综述侧重于对实验结果的描述，缺乏深入的思考和总结。建议作者进一步思考行为研究和认知神经机制研究之间的内在联系，阐述 2、3 和 4 这三部分内容之间的联系。

回应：非常感谢审稿人的意见。为了进一步增强当前综述内部结构的紧密性，深化行为研究和认知神经机制研究之间的内部联系。我们对正文做了如下修改：

第一，增强了第 2 部分有关“声学特征再现与发声器官运动复制共同表征嗓音模仿”的总结，即加入了嗓音模仿具有自发性和目标性的特征以及嗓音模仿涉及的两种自发性机制的区别（相关补充内容在 p5，第 2 段）：“在跟读范式的模仿条件下，说话人被要求模仿目标说话人的嗓音，这揭示了嗓音模仿也可以是针对特定目标声音而进行的产生。嗓音模仿所涉及的两种自发性也存在区别：在刺激-反应一致性范式中，自发性主要表现为基于感觉-动作联结的自动化反应。这意味着模仿行为几乎是一种由刺激触发的自动反应。与之不同，跟读范式中的自发性更侧重于无明确意图下的声学特性再现。也就是说，即使没有明确的模仿意图，说话人仍然能准确地再现目标说话人的声学特性。”

第二，在第 3 部分，我们增加了一段来具体描述嗓音模仿中的不同特征与其背后神经机制之间的关系，相关补充内容在 p8，第 2 段：“如前文所述，嗓音模仿具有两个核心特性：自发性和目标性。自发性可以进一步细分为两个层面：一是基于感觉-动作联结的自动化反应（通常观察于刺激-反应一致性范式中），尽管以往的成像研究并未直接针对使用刺激-反应一致性范式来研究嗓音模仿中自动化加工特性的相关神经机制，但手势动作模仿的成像证据仍可提供有用的参考，原因在于这些手势动作模仿的研究与刺激-反应一致性范式都是基于 ASL 理论模型的假设(Cracco et al., 2018)。这些研究发现观察和执行动作的过程会激活额下回和初级运动皮层，这些区域都与镜像神经元系统有关(Cracco et al., 2018)。镜像神经元系统，尤其是在猕猴脑中的前运动皮层的 F5 区域，被认为是模仿和语言发展的神经基础(Nguyen & Delvaux, 2015)。这一系统通过促进观察到的动作和声音的内部映射，为一般性的模仿行为提供了神经基础。在嗓音模仿中，这些镜像神经元就会启动。它们不仅帮助说话人准确地“听”到目标说话人嗓音的特点，还将这些信息转换为具体的发声指令，好让说话人的喉咙和嘴巴知道要怎么动才能模仿出相同的声音。另一是即便在无明确意图的情况下仍能准确再现声学特性（主要基于跟读范式的研究结果）。在目标性方面，模仿行为不仅是一种无意识的反应，也是一个有目标的过程。在模仿过程中，说话人通过逆向模型生成实现预定目标状态所需的动作指令。根据当前关于嗓音模仿的神经机制研究，我们发现无论是有意图的模仿还是无意图的模仿，这两种不同类型的模仿涉及的脑区都是相关的。这表明有意图模仿和无意图模仿在神经层面上可能共享相似的处理路径或网络。相应的行为证据发现，与非模仿条件相比，两种模仿方式都能导致声学特性的聚合。这意味着声学特性的再现可以是无意图、自发产生的，也可以是有意图、目标导向的。这两者的主要差异可能体现在声学特性再现的程度上。例如在无意图的模仿中，说话人可能会在快速的言语交流环境中与目标说话人在声学特性上逐渐接近和靠拢，以促进更有效的沟通和合作。而在有意图的模仿中，说话人可能会更加精细地调整，以消除自己与目标说话人在声学特性上的差异，从而更接近“声学特征完全相同”的目标。”

第四，在第 4 部分，即有关个体差异的结尾处，我们基于行为和神经实证研究所梳理的证据来对嗓音模仿的不同认知阶段进行了解释，这样做的目的一方面可以深化行为和神经之间的关系，另一方面使得个体差异部分的结论不那么宽泛（相关内容增加在 p11，第 2 段）：“综上，嗓音感知，感知映射产出，以及嗓音产出这三个主要加工阶段确实会影响嗓音模仿的行为后果。首先，在感知阶段，言语感知－产出整合模型与 ASL 模型都强调说话人需要准确地提取目标说话人有用的声学信息和可见的发声器官信息（如口型）。这些信息会影响后续的产出指令或发声动作。更准确地提取与目标嗓音相关的信息将有助于嗓音模仿的准确性。此外，这一过程与次级听觉皮层（如颞上回）有密切的关联。其次，在感知映射产出阶段，言语感知–产出整合模型指出，唱音障碍可能源于前向模型在预测发声器官的动作指令时存在的偏差。从 ASL 模型的角度来看，这种障碍也可能受到先前感觉-动作联结经验的影响。具体地说，不熟悉的听觉事件可能会受到已有感觉-动作联结经验的阻碍，而熟悉的听觉事件则可能受到这种经验的促进。这一认知过程与弓形束和基底神经节的活动有关。最后，言语感知–产出整合模型与 ASL 模型都强调了发声器官运动的灵活性。前者认为，为了满足前向模型预测的发声动作指令，需要足够灵活的发声器官运动。后者则指出，只有具备这种灵活性，才能丰富和扩展已有的感觉-动作联结经验。这一过程与初级运动皮层（如喉部，唇部以及舌部）的活动密切相关。通过这三个阶段的综合分析，不仅可以更全面地理解嗓音模仿的复杂性，还可以清晰地看到各种认知模型和神经基础如何共同作用于嗓音模仿。”

意见 11：不通顺的句子，例如“会激活颞上沟（superior temporal sulcus，STS）、颞上回（Superior Temporal Gyrus，STG）和颞中回（middle temporal gyri，MTG），这些与嗓音感知相关的脑区

(Frühholz & Schweinberger, 2021)。”

回应：感谢审稿人的意见。我们已经通读了修改后的全文，并调整了相应不通顺的句子（例如，p6，第 3 部分第 2 段）：“结果发现，与跟读单一目标说话人相比，当说话人跟读多名目标说话人时，会激活颞上沟（superior temporal sulcus, STS）、颞上回（Superior Temporal Gyrus, STG）和颞中回（middle temporal gyri, MTG）等与嗓音感知相关的脑区(Frühholz & Schweinberger, 2021)。”

第二轮

审稿人 1 意见：作者回答了我提出的问题。

审稿人 2 意见：作者根据审稿意见进行了认真和全面的修改，修改后的文章理论深度有明显提高，建议接受发表。

编委 1 意见：可以发表。

编委 2 意见：同意发表。

主编意见：稿件经过多位专家的审阅，作者进行了认真修改，达到发表水平，同意发表。