心理科学进展 2024, Vol. 32, No. 3, 499–513
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2024.00499

嗓音模仿认知神经加工的多阶段模型*

胡砚冰[1] 蒋晓鸣[1,2]

([1]上海外国语大学语言研究院; [2]上海外国语大学语言科学与多语智能应用重点实验室, 上海 201620)

摘 要 嗓音模仿是言语交流中关键的认知过程, 即对话一方(说话人)将感知到的另一方(目标说话人)的嗓音信号映射为自己的发声器官运动表征, 达到发声器官运动表征复制和目标说话人嗓音再现的目的。成像结果表明, 嗓音模仿的认知加工涉及颞上回到左侧额下回, 再到发声相关初级运动皮层的神经网络, 并且基底神经节在该网络中发挥协调作用。嗓音辨别能力、嗓音信号至发声运动表征的映射能力以及发声器官肌肉的控制能力的个体差异都会影响嗓音模仿的认知加工。未来研究应该考虑将嗓音模仿与发声障碍以及侵入电极技术结合起来, 旨在共同揭示脑与行为的因果机制, 并进一步应用于言语的终身发展、认知可塑性以及言语预期领域。

关键词 嗓音模仿; 嗓音再现; 发声器官运动复制; 模仿神经网络; 个体差异

分类号 B842

1 引言

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的多模态信息, 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。

言语感知−产出整合模型(speech perception-production integrational model)认为说话人会模仿目标说话人的言语信息, 将目标说话人的感知表征映射为自己的产出指令, 并使用这种产出指令来引导其接近目标说话人的言语信息(Gambi et al., 2022; Gandolfi et al., 2022; Pickering & Gambi, 2018; Pickering & Garrod, 2004, 2013)。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

收稿日期: 2023-08-04

* 国家自然科学基金委面上项目(31971037)、上海市教育发展基金会和上海市教育委员会“曙光计划”项目(20SG31)、上海市自然科学基金面上项目(22ZR1460200)支持。

通信作者: 蒋晓鸣, E-mail: xiaoming.jiang@shisu.edu.cn

[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。

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)。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。最后, 在嗓音层面, 说话人也会借助目标说话人的嗓音特征来调整自己的产出系统, 以便更准确地模仿目标说话人的声音(Pardo et al., 2013)。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。此外, 该模型还进一步阐释了在不同物种, 不同种族以及不同年龄段中声学特征相似性的出现, 如有研究发现亚洲象和海豹都可以模仿人类的嗓音, 具体表现这些动物在发声行为中表现出与人类相似的共振峰大小(Stansbury & Janik, 2019; Stoeger et al., 2012), 有研究发现嗓音模仿可以帮助说话人理解外种族口音(Adank et al., 2010), 还有研究发现 5 岁左右的儿童模仿妈妈的声音, 可以促进其与妈妈大脑的同步(Lin et al., 2023)。这些证据表明, 嗓音模仿在语言发展, 特别是二语习得中起着关键作用。然而, 为什么嗓音模仿有助于语言的发展和习得, 需要当前综述梳理嗓音模仿背后的认知加工机制。更重要的是, 在现实生活中, 随着人工智能语音技术的发展, 人类与机器的语音交互日益成熟。然而, 要使机器更好地模仿人类嗓音, 并让其听起来与人类的嗓音更为相似, 以实现更加人性化的听觉体验, 仍然面临着挑战(Cohn et al., 2022; Zellou et al., 2021)。

与传统观点中将说话人的感知和产出视为两个独立模块的模块化言语交流理论不同, 言语整合−产出模型提出, 说话人在感知目标说话人时, 同时也在调整自己的发声指令。这一模型强调了感知和产出实际上是相互依赖的认知过程(Pickering & Garrod, 2013)。这为解释说话人和目标说话人在嗓音声学特征上的相似性提供了理论基础。值得注意的是, 在日常言语交流中, 不仅有听觉信息, 还有视觉上的发声线索, 如口型。这些视觉线索也能影响模仿的认知过程, 意味着嗓音模仿也可以体现在说话人和目标说话人口型的相似性上。关联序列学习模型(associative sequence learning, ASL)认为嗓音模仿是特定效应器官感觉−运动联结学习的产物, 模仿过程涉及观察和执行相同动作的联结经验。具体来说, 在模仿过程中, 说话人首先观察目标说话人某一特定效应器官的行为(如嘴唇张开或闭合), 然后尝试使用对应的效应器官执行相同的行为, 从而实现对行为序列的学习(Heyes, 2001, 2011; Wu et al., 2019)。与此相关的是逆向模型(inverse model)的参与(Belyk et al., 2016)。逆向模型主要关注如何将感知到的目标说话人的嗓音信号转化为说话人相应的发声运动指令(Chen et al., 2021)。与前向模型(即在说话前预先设定嘴唇、舌头和其他发声器官的运动以产生期望的嗓音效果)不同, 逆向模型更注重目标导向的映射过程。具体来说, 这意味说话人的发声运动指令是受到目标说话人发声动作的引导, 以便与目标说话人在特定特征上(例如, 发声器官的运动特点)达到高度匹配。

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 et al., 2019; Belyk, Brown, et al., 2021; 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。

本文基于说话人角度[2], 分别从三个部分来解决这些问题: 第一部分主要基于嗓音模仿的两种模型来梳理与其相关的嗓音模仿认知加工阶段, 并阐述嗓音模仿背后认知加工过程的特异性; 第二部分通过梳理与嗓音模仿范式密切相关的成像研究来回答嗓音模仿背后涉及的神经网络, 解释认知加工特异路径背后涉及的脑机制; 第三部分从个体嗓音辨别能力、嗓音感知映射发声器官运动指令的能力以及发声效应器官控制能力的个体差异角度, 来探究这些因素如何影响嗓音模仿。

2　声学特征再现与发声器官运动复制共同表征嗓音模仿

言语感知−产出整合模型和关联序列学习模型分别从嗓音声学特征和发声器官运动表征角度来阐明嗓音模仿的认知过程。前者认为嗓音模仿的认知过程需要说话人再现(reenactment)目标说话人嗓音信息的认知过程参与(Mercado et al., 2014)。后者从发声器官运动表征角度认为嗓音模仿涉及说话人观察目标说话人发声器官的运动表征, 然后使用相应的效应器官复制(copy)同样的动作(Cracco et al., 2018)。综合两种解释可以发现, “再现”和“复制”都意味着说话人感知目标说话人的某个特征, 其特征可以是嗓音中的声学特征, 也可以是发声器官运动的特征。在此基础上, 说话人执行了一个“相同”的特征。更重要的是, 这两种解释不仅强调了嗓音模仿中的感知和发声过程, 而且表征了感知映射发声的认知过程。然而, 哪些指标可以测量嗓音模仿的认知过程, 需要进一步梳理。

基于嗓音再现的解释认为, 嗓音模仿具体表现为当被试接触(vs.无接触)目标说话人声音之后所产生的声音与目标说话人声音更具有相似性(Goldinger, 1998)。Goldinger (1998)首次采用跟读范式(shadow paradigm)来研究基于此定义下的嗓音模仿认知过程。具体来说, 该实验涉及两组被试: 说话人组和听话者组。说话人被要求收听目标说话人所产生的声音, 声音结束后, 说话人被要求重复这些声音。之后, 要求听者组进行 AXB 任务, 在这一任务中会向听者依次呈现三个听觉刺激(A、X、B)。其中, 听觉刺激(X)是之前记录目标说话人所产生的声音, 而 A 和 B 是由说话人组产生与目标说话人同一言语内容的声音刺激, 其中 A 是之前任务中说话人跟读目标说话人所产

[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。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 200 ms, 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 350 ms)。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。

生的声音刺激(跟读条件), B 是作为基线条件的声音刺激(即没有接触目标说话人之前, 说话人录制好的录音)。结果发现, 相比于基线条件, 听话者组被试认为跟读条件的声音与目标说话人声音更加相似(Goldinger, 1998; Pardo et al., 2013; Pardo et al., 2017; Pardo & Remez, 2021)。这项研究表明, 当任务没有要求被试模仿目标说话人声音时, 被试依然会模仿目标说话人的声音。

然而, Goldinger (1998)的研究存在这样一个问题, 对于跟读范式中被试模仿认知加工的测量依赖听者的主观判断, 即让新的一组被试(听话者组)直接对说话组在两个不同条件(跟读条件 vs.基线条件)所发出的声音与目标说话人录音的相似性进行比较。在这些研究中, 依然不清楚, 说话组被试在跟读范式中模仿了什么? 基于这一问题, 有研究同样采用跟读范式, 但是采用了有关模仿加工的不同测量方法。这些研究的基本逻辑是, 测量目标说话人组和跟读条件下的说话人组的声音是否发生声学聚合(acoustic convergence) (Garnier et al., 2013; Pardo et al., 2013; Pardo et al., 2017)。统计思路是, 首先测量目标说话人、说话人组在基线条件和跟读条件下所产生声音的声学参数(如基频、元音共振峰等), 然后将这些声学参数转换为欧氏距离差分数(difference−in−distance scores, DID), 用以评估声学聚合(Pardo et al., 2017)。计算出两种类型的 DID。第一种类型 DID 是通过比较基线条件与目标说话人之间的声学欧氏距离的差异(baseline-model), 第二种类型 DID 是通过比较跟读条件与目标说话人之间的声学欧氏距离的差异(shadow-model)。结果发现, 第二种类型的 DID 显著小于第一种 DID, 这表明, 说话人组被试在跟读条件的录音与目标说话人的录音发生了声学聚合(Pardo et al., 2017)。

然而, 这种跟读条件的声学聚合在多大程度上可以解释模仿的认知过程呢? 有研究将跟读任务中观察到的声学聚合效应与任务要求被试模仿所产生的声音进行比较, 以进一步说明这种声音聚合可以解释为模仿的认知加工(Dufour & Nguyen, 2013)。这项研究选取 22 个以/e/结尾和 22 个以/ɛ/结尾的双音节单词。在目标任务中, 双音节单词以听觉形式通过耳机进行呈现, 其中要求一组被试自然清晰地跟读目标说话人录音中的这些单词(跟读组), 要求另一组被试模仿目标说话人的具体发音(模仿组)。任务前后分别进行了一项测试, 把这些双音节单词以小写字母形式视觉呈现在屏幕中央, 要求被试大声读出这些单词。为了排除前测带来的练习效应, 另外 44 个双音节单词, 其中一半以/e/结尾, 另一半以/ɛ/结尾, 只在目标任务和后测中使用。并且将在前测中出现的单词设定为基线词(baseline word), 将没有在前测出现的单词设定为新词(new word)。结果发现在测试和后测阶段, 跟读组和模仿组都出现了相同的结果模式, 即新词中的/e/和/ɛ/的第一共振峰上都出现了显著差异。这项研究结果表明, 模仿组和跟读组都发生了声学上的聚合效应[3], 并说明这种聚合效应可以解释为一种模仿的认知过程。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, 不管是任务相关还是任务无关的模仿, 说话人都会在一定程度上模仿目标说话人的声音。

基于发声运动表征复制的观点认为, 说话人与目标说话人发声器官运动表征不一致(vs.一致)时, 会诱发其更长的发声时间延迟, 即与目标说话人一致的运动表征可以促进说话人发声运动表征复制的表现, 与目标说话人不一致的运动表征

[3]该名目标说话人具有标准的发音, 可以从声学特征显著的区分/e/和/ɛ/, 然而招募的被试都含有一定的口音, 不能在声学特征层面区分/e/和/ɛ/, 为此跟读组和模仿组被试之所以可以区分/e/和/ɛ/, 是因为与目标说话人声学特征发生了聚合。

则会干扰说话人发声运动表征复制的表现(Wilt et al., 2023)。有研究采用刺激反应一致性范式(stimulus−response compatibility paradigm)解释这种定义, 该研究通过呈现听觉刺激(/ba/或/da/), 并且与被试的发声任务要求形成一致或不一致的发声条件, 进而影响了被试的发声延迟。被试的任务是忽略耳机中的听觉刺激, 产出视觉提示所呈现的音节(/ba/或/da/), 为此就会操纵刺激感知与发声器官产出的一致性条件(如一致条件: /ba/−/ba/, 不一致条件: /ba/−/da/)。结果发现, 不一致条件的发声延迟时间显著长于一致条件(Galantucci et al., 2009)。这表明当个体感知到的刺激与实际产出刺激有冲突时, 会影响个体相应发声效应器官的动作表征, 这种影响表现在发声延迟时间上。然而, 这种刺激−反应一致性效应是否具有反应效应器官的特异性呢? 比如, 仅当利用发声相关的效应器官做反应时才会出现一致性效应, 利用别的效应器官做反应则不会出现一致性效应。这一问题对应了 ASL 对模仿认知过程的解释, 即模仿过程涉及观察和执行相同效应器官的动作。为了解决这个问题, Galantucci 等(2009)进行了第二个实验, 即被试除了发声任务, 还需要进行一个按键任务, 即忽略听觉刺激, 根据视觉提示通过肢体运动效应器官(即手指)进行按键。结果发现, 只有当被试进行言语产出任务时, 才会出现一致性效应。结果表明当说话人感知某一听觉刺激时, 会快速通过人声感知−运动映射机制形成一种与发声效应器官对应的产出表征, 当实际言语产出任务与这种产出表征一致时, 则会促进相应的产出行为, 当与产出表征不一致时, 则会抑制相应的产出行为。基于目前证据表明, 这种由刺激−反应范式诱发的一致性效应涉及说话人与目标说话人在相同效应器官上感知映射运动表征的过程, 如感知手的运动只会影响相应手的运动表征(Heyes, 2011), 感知嘴部的运动只会影响嘴的运动表征(Virhia et al., 2019)。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。

综上所述, 跟读范式中的跟读条件和刺激反应一致性范式的任务都不要求说话人对目标说话人嗓音进行模仿, 进而通过行为反应指标来测量模仿的认知加工过程, 这表明嗓音模仿可以是自发产生的。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。此外, 两者在测量模仿认知加工的指标上也存在差异, 跟读范式对于模仿的测量体现在说话人与目标说话人声音特征间存在相似性, 或者两者的声学特征发生聚合, 这一指标反映了嗓音模仿发声这一阶段; 刺激−反应一致性范式对于嗓音模仿测量体现在嗓音感知映射发声器官运动表征和实际发声动作之间的一致性, 这一指标包含了嗓音模仿感知和感知映射发声器官运动表征这两个加工阶段。这些证据表明嗓音声学特征再现和发声器官复制的测量指标共同表征了嗓音模仿的认知过程。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。

3　嗓音模仿的特异路径: 认知神经多阶段加工模型

嗓音模仿包含感知、感知映射发声运动表征以及发声产出三个加工阶段, 这并不意味着嗓音模仿等于三个加工阶段的“和”。相比于独立的三

个加工阶段，嗓音模仿认知加工涉及整合和协调感知−发声运动认知阶段所涉及的神经网络。当前部分通过梳理与嗓音模仿范式相关的成像研究来澄清其模仿过程背后的特异性神经机制。此外，嗓音模仿除了在行为指标上存在特异性(即说话人对目标说话人相应发声效应器官的复制和再现目标说话人嗓音)，嗓音模仿与感知−发声运动肯定在神经机制上也存在差异。当前部分将梳理嗓音模仿相关的成像研究，并与经典感知−发声运动神经网络进行对比，进而阐明嗓音模仿与感知−发声运动加工在神经机制上的联系和区别。

一项功能性磁共振成像(functional magnetic resonance imaging, fMRI)研究要求说话人跟读多名或一名目标说话人提前录制好的双音节假词[4]，使用假词的目的是为了防止说话人对刺激中的语义信息进行加工，实验者使用事件相关的序列采样设计扫描了被试跟读过程中的脑区活动变化，并且记录了被试产出声音中的声学特征(如 F0，时长等) (Peschke et al., 2009)。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)。此外，还激活了额下回(inferior frontal)，初级运动皮层(primary motor cortex, M1)这些与言语发声计划或运动相关的脑区(Pisanski et al., 2016)。随后 Peschke 等(2009)进行了个体差异分析，结果发现更加延迟的言语产生反应[5]与左侧顶叶盖(left parietal operculum, LPO)区域的激活成正相关。以往研究表明，LPO 同时参与听觉感知与产出运动的认知加工，并且是参与听觉感知映射发声器官运动过程中的重要脑区之一(Hickok & Poeppel, 2000)。为此，Peschke 等(2009)在个体差异上的结果可能说明了，听觉感知至言语产生映射之间认知过程的整体难度与复杂性与 LPO 的激活密切相关。然而，这项研究的结果在多大程度上可以解释为模仿的神经机制呢？

一项 fMRI 研究让被试进行了 5 项任务，其中两项是基线任务：(1)在感知参照任务中，要求被试被动听元音；(2)在产出参照任务中，要求被试产出屏幕上呈现的元音。另外三项为当前研究感兴趣的三项任务：(3)在跟读产出任务中，要求被试产出耳机中呈现的元音(任务无关的模仿认知加工[6])；(4)在元音模仿任务中，要求被试模仿录音中的声音(任务相关的模仿认知加工[7])；(5)在抑制产出任务中(抑制模仿的认知过程)，提前告知被试声学聚合的现象，然后要求被试忽略录音中的声学线索，用自己的发声方式进行产出，除了扫描这些任务的认知过程外，还扫描了被试的静息态数据。此外，在其中三项感兴趣的任务中，被试需要额外完成 Go−No Go 任务[8]，即看到绿色的注视点时才能进行相应的产出任务，看到红色的注视点时不能进行相应的产出任务(Garnier et al., 2013)。结果发现，任务无关模仿和任务相关模仿都诱发了相同的嗓音感知−发声运动网络，与以往研究发现的嗓音感知−发声运动网络相似，这项研究发现的共享感知网络包括双侧 STG 的激活，延伸到罗兰氏叶盖(Roland Operculum)和左侧脑岛(insula)，同时还有左外侧额下回，特别是布洛卡区的额下回(Inferior Frontal Gyrus, IFG)三角部和前额区 BA8 以及双侧顶下小叶(Inferior Parietal Lobe, IPL)区域，双侧缘上回(bilateral supramarginal gyrus, SMG)与右侧角回(angular gyrus)，在边缘系统中还发现了右侧丘脑(thalamus)和左后侧扣带回皮层(cingulate cortex)的共同激活；这项研究发

[4] 假词也被称为伪词或非词，是指那些在语音和形式上看起来像真实的词语，但实际上并没有任何已知的含义或语义的词语。

[5] 为了排除正确率−反应时平衡带来的影响，即反应时长可能是因为任务难度所造成的，Peschke 等(2009)进行了额外分析，将正确与错误试次的产出反应时进行对比，结果显示正确试次的产出反应时长于错误试次的产出反应时(t = –2.39; df = 19; p = 0.0276; 正确试次的产出反应时 = 488 ms, 错误试次的产出反应时= 469 ms)。如果任务困难导致产出反应时的增加，那么这些困难并没有导致错误试次的产出反应时增加，故排除正确率−反应时平衡带来的影响。

[6] 任务无关的模仿认知加工是指在该任务指令中不涉及明确要求被试去复制，再现或者模仿目标声音。

[7] 任务相关的模仿认知加工是指在该任务指令中明确要求被试去复制，再现或者模仿目标声音。

[8] 此任务的优势在于进行感知−产出的双任务时，可以探究“主动感知”的认知过程，如元音模仿任务中 No Go 试次的 BOLD 信号与其静息态进行对比。也可以仅探究产出的认知过程，如元音模仿任务中 Go 试次的 BOLD 信号与元音模仿任务中 No Go 试次进行对比。

现的共享产出网络包括, M1 和运动感觉皮层的双侧激活, 延伸到 IFG 三角部和 SMG, 这一产出网络涉及 STG 中的初级听觉皮层, 延伸到罗兰氏叶盖和脑岛, 在顶叶后部区域, 包括楔前叶(precuneus)和整合皮层(associative cortex)以及边缘系统(前扣带回、丘脑)、小脑、壳核(putamen)、红核(red nucleus)和右侧基底节(right basal ganglia, BG)发现进一步的共同激活(Garnier et al., 2013)。这项研究提取了感兴趣任务(即任务 3, 任务 4 以及任务 5)产出后的声学特征(F0), 并用录音中目标说话人的 F0 与感兴趣任务中说话人产出的 F0 做相关。结果发现, 任务相关模仿的斜率大于任务无关模仿, 任务相关和任务无关模仿的斜率都大于抑制模仿。这表明, 基于相关性斜率的指标可以表征说话者模仿的程度。研究者进一步将斜率与三种模仿认知加工中, 激活的感知网络和产出网络中的脑区激活程度分别做相关, 结果仅在与听觉感知网络相关的双侧听觉皮层、双侧 SMG 和左侧韦尔尼克区存在显著相关性(Garnier et al., 2013)。这些结果表明, 任务无关和任务相关的模仿过程都涉及大脑背侧感觉−运动网络的参与, 对于听觉表征映射到发音表征的认知过程非常重要(Hickok & Poeppel, 2000, 2004)。这一认知网络在模仿任务中得到了实证验证, 进一步强调了它在嗓音模仿加工方面与单纯嗓音感知和产出的认知过程有明显的区别和特异性。更为详细地说, 在模仿加工的关键步骤中, 说话人首先需准确地感知目标说话人的声音特性, 然后将这些感知到的声学信息转化为自己发声器官的运动指令。通过执行这些运动指令, 说话人能够用自己的声音复制目标说话人的声音特性。这一流程揭示了嗓音感知与发声动作在嗓音模仿认知加工中具有至关重要的协同作用。这引出一个问题: 与这种协同认知加工紧密相关的神经机制是什么?

一项 fMRI 研究要求被试分别进行三项任务[9]: (1)在音高模仿任务中, 要求被试模仿刚才耳机呈现的 4 个不同韵律的音符; (2)在非模仿发声任务中, 要求被试根据视觉所提示的熟悉旋律名称, 发出该韵律中的前 4 个音符; (3)在音高辨别任务中, 要求被试通过按键来表明最后一个音符和前 3 个音符是否相同。联合分析的结果发现, 相对于音高辨别任务和非模仿发声任务, 模仿任务中更多的激活了壳核、SMA 和口部感觉运动皮层(Belyk et al., 2016)。壳核作为 BG 中的一部分, 具有执行动作选择, 习得新运动序列以及执行动作调节方面的功能(Shmuelof & Krakauer, 2011)。这可能表明 BG 在嗓音模仿过程中的重要性, 原因在于嗓音模仿涉及再现日常中不经常接触的声音, 如外种族口音(Adank et al., 2010), 异性的声音(Cartei et al., 2020), 无语义信息的音节(Pardo et al., 2013)。在此结果基础上, Belyk 等(2016)提出了嗓音模仿认知神经模型(neural model of vocal imitation), 这一认知神经模型涉及 STG 的后部, 并沿着弓状束(arcuate fasciculus, AF)传送到额叶, 再由 IFG 投射到 M1, 初级皮层执行运动指令以再现目标声音。重要的是, Belyk 等(2016)研究发现 BG 相关的皮质环路参与听觉目标映射运动指令的认知过程。然而, Belyk 等(2016)在模型中对于 AF 的假设是基于以往关于言语感知−产生的文献, 仅凭当前成像结果很难给出相应的证据。一项 fMRI 研究采用扩散加权成像(Diffusion−Weighted Imaging, DWI)的技术来研究 AF 在无语义(即伪词产出任务)和有语义任务(即真词产出任务)中 AF 的子功能结构, 结果发现, 在无语义任务中, AF 在 STG 与 IFG 之间起到桥梁作用, 而在有语义任务中, AF 在 MTG 与 IFG 之间起到桥梁作用(Janssen et al., 2023)。这表明, STG−AF−IFG 起到嗓音感知映射发声动作指令的作用。这与 Belyk 等(2016)在音高模仿任务中提及的模型是一致的。为了进一步验证这一模型, 研究者(Belyk, Brown et al., 2021)采用了更高空间分辨率的 7T fMRI 进行了研究, 研究范式采用了 Belyk 等(2016)中的音高模仿任务, 但是分别要求被试用吹口哨和唱歌的方式进行产出, 通过对比嗓音模仿(包含两种产出方式)与静息态的成像数据, 结果发现, 位于 M1 的腹侧和背侧喉部运动皮层(ventral/dorsal larynx motor areas, v/dLMC)这些与产出相关的脑区激活以及 STG, 内侧膝状体(geniculate nucleus of the thalamus)这

[9] 这项研究的目的旨在表明嗓音模仿除了感知和产出两个基本的认知加工阶段, 还包括逆向模型的参与。为此, 研究者将音高模仿任务(涉及嗓音模仿认知加工)所激活的脑区, 分别与非模仿发声任务(涉及嗓音产出认知加工)和音高辨别任务(涉及嗓音感知认知加工)中所激活的脑区进行对比, 并进一步将对比的结果进行联合分析, 联合分析的结果表明了逆向模型的神经机制。

些与听觉感知反馈相关脑区的激活。

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。

通过综合直接研究嗓音模仿认知加工的成像研究，我们不仅理清了与感知−发声运动密切相关的神经通路，如 STG−AF−IFG−M1，还发现了皮质下通路，包括纹状体(BG)和内侧膝状体，这些区域与根据目标嗓音进行运动序列的选择、协调与执行有着密切关系(如图 1 所示)。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，

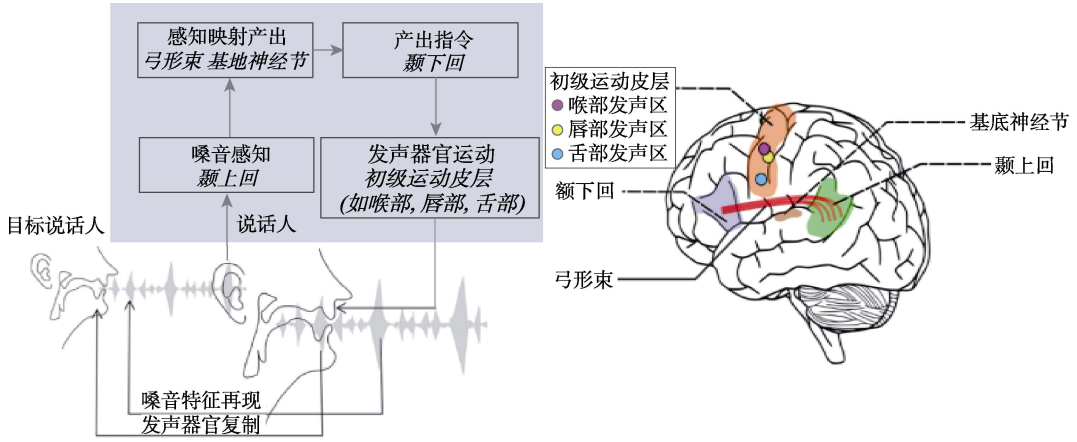

图 1 嗓音模仿认知神经多阶段加工模型

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。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。

4　嗓音感知、感知映射产出以及嗓音产出能力影响嗓音模仿

如前所述, 嗓音模仿认知加工中关键的认知过程在于将目标说话人的感知表征映射到说话人自己的发声器官运动表征。这表明嗓音模仿的认知过程包括了听觉感知、说话人嗓音感知映射产出表征以及嗓音产出三个主要阶段。基于此, 当前部分将详细梳理与这三个主要加工阶段密切相关的个体差异能力如何影响说话人模仿目标声音的程度(Kim & Clayards, 2019)。通过梳理嗓音模仿认知加工相关的个体差异, 有利于为一些特殊群体提供精准的干预方案, 比如唱音障碍等发声障碍群体, 也有利于对言语习得提供新的视角。

一项研究招募了一名母语为英语的女性目标说话人, 要求其重复多次录制 head 和 had 两个单词。随后挑选出发音, 质量最好的两个音频, 其中一个音频为 head 的发音, 另一个音频为 had, 再通过 TANDEM-STRAIGHT 软件对 head 和 had 进行声音变形(speech morphing), 生成 20 条连续声音固定步长梯度变化的音频, 从/hɛd/到/hæd/过渡。然后由 5 名母语为英语的听者对这 20 条音频进行迫选任务, 即判断音频属于 head 还是 had, 从 50%选择 had 的行为反应中筛选具有歧义的音频, 即听者无法准确判断这条音频属于 head 还是 had 范畴, 进一步基于两条具有明显 head 和 had 范畴的音频和一条歧义音频, 再通过第一共振峰(F1)和第二共振峰(F2)的值从 17 条音频中选择 4 条, 最后组成 7 条梯度分布较好的音频, 即从 head 频谱特征(即 F1 和 F2)固定步长过渡到 had, 在此基础上, 每个音频又创建步长为 40 ms, 100 ~ 340 ms 范围的 7 条基于时长梯度变化的音频, 综上共生成 49 条在元音频谱(即 F1 和 F2)和时长两种维度连续体的音频, 作为后续感知任务的材料。对于产出任务的材料, 同样是基于元音频谱和时长两种维度, 与感知任务材料不同的是, 对于频谱维度, 选择了两个极端具有明显范畴的音频和 1 条歧义音频, 即一共 3 条音频, 对于时长维度创建了步长为 80 ms, 60~380 ms 范围的 5 条基于时长连续体的音频, 最终生成了 15 条音频。在感知任务中, 要求说话人对音频是属于 head 还是 had 做迫选任务。在产出任务中, 说话人首先进行基线产出任务(baseline task), 说话人在这项任务中被要求清晰且自然地读出电脑屏幕视觉呈现的单字, 其次说话人进行模仿产出任务(imitation task), 说话人在这项任务中被明确要求在听到模仿目标(target)后尽可能地模仿目标刺激, 两项产出任务都记录了相应的声学数据。在感知任务的数据分析中, 基于频谱和时长维度, 分别拟合每名说话人的逻辑回归, 并提取每个模型的系数作为不同维度的感知权重, 即代表每名说话人利用音频中声学信息的能力, 在产出任务的数据分析中, 最重要的是对模仿表现指数的操纵性定义, 即基线条件的声学参数[10]与目标声学参数之间的欧式距离减去模仿条件与目标声学参数之间的欧式距离(|Target-Baseline| − |Target-Imitation|), 如果模仿表现指数为负, 表明说话人与目标说话人嗓音趋异, 如果模仿表现指数为正, 表明说话人与目标说话人嗓音趋同。结果发现, 被试在元音频谱维度的感知权重可以正向显著预期其在时长这一声学参数的模仿表现(Kim & Clayards, 2019)。这项研究运用心理物理学原理来探究嗓音辨别能力对于嗓音模仿的影响, 通过操纵刺激声学特征连续梯度的变化, 进而测量说话人在嗓音辨别和嗓音模仿任务中的行为表现, 从结果可以看出这些测量指标对于相应操纵是敏感的, 即被试在感知任务中利用声音中的频谱信息越多, 越有利于其嗓音模仿。这些发现进一步表明, 基于嗓音感知的嗓音辨别能力可以有效的预测嗓音模仿能力。

此外, 如果说话人对目标说话人的感知表征映射到自己产出表征这一过程存在加工障碍, 会同样影响模仿的准确性。研究发现, 对于一些唱音障碍(poor-pitch singing)群体, 其感知和发声运

[10] 这里声学参数为 F1(Hz), F2(Hz), 时长(ms)。

动系统都正常，然而这些唱音障碍群体模仿的音高与目标音高总是在听感[11]上相差很大，这可能是因为音高感知到发声器官运动映射的过程存在认知障碍(Belyk, Johnson & Kotz, 2018; Belyk, Lee & Brown, 2018)。根据这些基本假设，有研究要求说话人进行两个阶段的任务，在第一阶段任务中，要求说话人模仿 4 个目标声音，并将说话人的录音记录下来。在第二阶段任务中，说话人模仿两种类型的目标声音，第一种是在第一阶段录制的自己的声音，第二种是录制陌生人的目标声音，然后基于音高，分别计算第二阶段被试模仿的声音与两种目标声音的差异分数。在分析中，根据说话人总体的模仿表现[12]分成模仿音高准确组和模仿音高不准确组，结果发现，模仿音高不准确组对自我目标声音的模仿表现要显著好于模仿陌生他人目标的声音，然而在模仿音高准确组中却没有发现两者存在差异(Pfordresher & Mantell, 2014)。这些结果表明，音高感知映射发声器官运动的认知过程存在障碍的原因在于，唱音障碍群体对于自己声音是非常熟悉的，所以当他们感知自我声音，进而映射到相应的产出指令，从而产生与自我目标声音相似的声音；但当唱音障碍群体感知陌生声音时，不能映射与陌生目标说话人声音对应的发声指令，为此，唱音障碍说话人模仿陌生声音时，在听感上就会明显区分其模仿自己的声音和陌生目标声音。相对于唱音障碍群体，正常群体不仅对自我目标声音可以映射到相应的发声指令，而且对陌生的目标声音同样可以正常映射到自己的发声指令，并且其模仿的声音和陌生目标声音在听感上是相似的。

最后，嗓音模仿的认知过程与说话人对于发声控制的能力密切相关，比如对于声带、喉部肌肉、上颚部、唇部等发声器官的控制，并且喉部、唇部以及舌部运动的神经表征都位于初级运动皮层(Belyk & Brown, 2017; Belyk & McGettigan, 2022; Brown et al., 2021; Kuhlen & Abdel Rahman, 2023)。皮质−延髓假说(corticobulbar hypothesis)认为，人类发声行为的灵活性，可归因于初级运动皮层到脑干运动神经元的直接连接，而在非人类灵长类动物中是间接的(Pisanski et al., 2016)。一项 fMRI 研究招募了歌手组和控制组作为实验被试，要求所有被试模仿 5 个声音刺激，在被试模仿的过程中记录声道和脑的成像，其中 4 个声音刺激都是通过改变其中一个正常声音的声学参数(即基频: F0 和声学意义的声道长度: vocal tract length, VTL)得到，具体来说，分别生成了低 F0 短 VTL，低 F0 长 VTL，高 F0 短 VTL，高 F0 低 VTL 四种声学刺激。声道成像结果发现，歌手(vs.控制组)对于不同声音刺激与其声道运动的幅度更加相关，比如对于长 VTL 的刺激，其对于声道的收缩越松，对于短 VTL 刺激，其对于声道的收缩越紧(Waters et al., 2021)。fMRI 扫描了被试在模仿准备和模仿执行阶段的全脑血氧信号，具体来说，模仿准备阶段涉及被试嗓音产出开始之前的认知加工阶段，模仿执行涉及被试正在嗓音模仿产出的认知加工阶段。该研究进一步采用表征相似性分析结合兴趣区探照灯分析(searchlight analysis)，该方法可以确定哪些脑区与特定的认知过程有关，具体来说，通过不同 VTL 刺激预先构建被试水平的理想不相似性矩阵[13]，并将基于模型模拟的数据映射到脑区的血氧信号数据，进而确定两者之间的相似性。结果发现，歌手(vs.控制组)在模仿准备阶段，显示出右中央沟/中央后回(right central sulcus/post-central gyrus)的重叠区域以及海马和丘脑这些脑区对于 VTL 相似性矩阵更强的表征，然而，类似的分析方法没有发现在模仿执行的认知加工阶段出现显著差异(Waters et al., 2021)。这些结果表明，具有声道控制训练经验的个体具有更强的模仿能力，并且在发声动作之前，就已经在将目标声音映射在自己的产出表征。

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

[11] 这里的听感一般来自于第三方被试的听觉感知。

[12] 模仿表现，即被试模仿产出的音高距离目标音高的距离，如果这一距离小于 50 音分则划分为模仿音高准确组，如果这一距离大于 100 音分则划分为模仿音高不准确组。

[13] 理想不相似性矩阵是指用于衡量理论模型预测的不同刺激之间的差异，这个矩阵代表了实验结果中期望看到的相似性模式，继而通过实验数据与理想不相似举证进行比较，进而可以评估数据中的模式是否与理论模型的预测相符。

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。

5　总结与展望

本文从言语中嗓音模仿的定义以及行为证据、嗓音模仿的神经机制、嗓音模仿的个体差异因素三方面来梳理其背后的认知神经基础。经过对聚焦于该问题的梳理, 本综述发现, 人类族群中的嗓音模仿不仅需要满足说话人与目标说话人嗓音中声学特征的相似(声学线索再现), 而且需要满足说话人发声器官与目标说话人发声器官运动表征的对应(发声器官复制)的条件。现有证据主要通过间接测量说话人与目标说话人声学参数欧式距离的大小, 来表征哪个声学参数发生了模仿, 比如元音共振峰(Pardo et al., 2017)、辅音声门开放时间(Yu et al., 2013)、基频等(Pardo et al., 2013)。其次, 通过梳理与嗓音模仿相关的神经通路, 发现脑左侧颞上回→弓形束→额下回→喉部运动皮层的神经通路以及基底神经节在其中的重要作用。此外, 嗓音模仿可以在任务无关条件下发生, 并且相对于领域普遍的的感知−发声运动加工, 嗓音模仿复制和再现的过程需要有目标性。最后, 通过梳理个体差异相关研究, 发现嗓音感知, 嗓音感知−映射发声器官的运动表征和发声产出三个方面能力的个体差异都会调节说话人在模仿目标说话人嗓音时, 其发声器官复制的能力和声学线索再现的能力。

尽管嗓音模仿认知加工中的一些问题已经得到解决, 但对该问题的研究尚处早期阶段, 有较多缺口需要未来研究探索。首先, 现有关于发声器官的模仿测量都是基于说话人在刺激−反应一致性范式中的产出延迟时间所定义, 而更加有说服力的方法应该直接测量发声器官的运动。其次, 目前已有理论表明, 嗓音模仿在言语交流起着非常重要的作用, 然而却鲜有实证证据的支撑。最后, 嗓音模仿相关范式可以应用到一些言语发声障碍群体的诊断筛查中, 原因在于嗓音模仿涉及对低级发声器官的操纵性定义。

5.1　嗓音模仿范式应用于脑机接口: 对于发声障碍的启示

有关嗓音模仿机制的研究在脑机接口(brain-computer interfaces, BCIs)中的应用尚处早期阶段。然而, 在一些发声困难群体(如构音障碍)的脑机接口应用面临着精准控制发声相关运动皮层的挑战。目前, 随着颅内皮层脑电图(intracranial ElectroCorticoGraphy, iECoG)技术的发展, 可以精准定位到与特定发声线索(如音调)相关的运动皮层(如喉部运动皮层), 旨在达到从神经信号中解码合成特定嗓音线索的目的(Liu et al., 2023)。然而, 这项研究存在两个局限性: 第一, 如本文所述, 产出不是独立于感知的认知过程, 感知依然对于产出的认知加工存在不可或缺的作用; 第二, 来自单一模态的信号对于解码嗓音线索肯定是存在局限性的, 比如解码更贴近自然的嗓音线索(Liu et al., 2023; Zhang et al., 2022)。

针对现有研究没有考虑感知在产出加工中的作用这一局限, 嗓音模仿范式和 iECoG 结合的未来研究可以很好地弥补这一缺陷。原因在于嗓音模仿认知加工涉及产出的嗓音线索可以最大限度地再现感知目标嗓音线索的认知过程, 即同时包含了感知与产出认知加工的参与。更重要的是, 相比于普通的产出任务, 基于嗓音模仿范式的产出任务, 可以激活更加精细的发声相关运动皮层(Belyk, Eichert et al., 2021; Bono et al., 2022)。iECoG 带来的关于神经空间和时间方面的优势, 同样可以深化嗓音模仿的认知神经机制。比如嗓音模仿涉及的 STG−AF−IFG−M1 神经通路, 其所

对应的时间特异性信息, 需要利用 iECoG 技术进行揭示。

此外, 单模态信号不足以解码更加复杂的嗓音信号这一局限。嗓音模仿范式中除了可以测量 iECoG, 还可以对被试的发声器官直接进行测量。比如电子声门仪(Electroglottograph)通过将两块电极放置甲状软骨两侧, 进而测量声门闭合和开启时的电流值, 以此来测量说话人声门运动的情况(Herbst, 2020), 超声舌部成像(Ultrasound Tongue Imaging)通过将超声探头放置在下巴下方, 进而捕捉舌头的运动和形状图像(Mousikou et al., 2021), 电磁发声造影技术(Electromagnetic Articulography)通过在舌头、唇部和颌骨等部位附加电磁传感器, 进而测量口腔和喉部运动(Paroni et al., 2021)。在此基础上, 可以提取在嗓音模仿范式下产出特定嗓音线索的同步神经、肌肉、舌部成像的多模态信号中来进一步解码合成嗓音信号。更重要的是, 如果一个人的发声器官受到损伤或者无法正常工作(即发声障碍群体), 他可能无法产出与目标嗓音相匹配的声音, 即使他能够准确地感知和分析目标嗓音。为此, 研究发声障碍群体的嗓音模仿认知加工, 有利于解答嗓音产出认知加工阶段在嗓音模仿认知加工中的因果机制。

5.2 嗓音模仿作用于言语的终生发展: 习得与老化

根据前文所述, 嗓音模仿可以自发产生, 并且其感知映射发声的认知过程具有目标性。在语言的发展和习得过程中, 嗓音模仿在很大程度上促进了儿童对音素、音调和韵律特征的习得。原因在于, 在早期言语发展阶段, 儿童会将感知到成人的嗓音发声特征(如发声效应器官运动特征或嗓音中超音段线索特征)作为一种目标, 这种目标会下达相应的指令来引导发声器官如何接近其目标的感知特征, 这一过程会帮助儿童逐渐形成对语言音素和语调的敏感性, 从而帮助儿童可以更好地理解言语规则(Cartei et al., 2020)。未来研究可以采用跟读范式中嗓音特征相似性指标来评估儿童言语习得能力, 具体来说, 可以通过比较儿童在跟读条件下产出的嗓音特征与标准发音之间的声学距离。

此外, 随着中国人口老龄化趋势日渐增高, 老年人口往往面临嗓音质量下降、发音准确性降低以及语言流利性减弱等问题, 并且进一步影响嗓音信号在其神经动态特征的同步性(Mai & Howell, 2023)。这可能导致他们在交流和表达中遇到困难, 影响日常生活质量和社交互动。最近一项研究表明, 有音乐(vs. 无音乐)训练[14]的老年人[15]在有噪音条件下, 对于音节的感知能力更强, 更重要的是, 这些经过音乐训练的老年人与无音乐训练的年轻人[16]在噪音条件下对音节的感知能力相当, 有趣的是, fMRI 结果发现, 经过音乐训练的老年人在感觉运动区(sensorimotor region)中保留了与无音乐训练的年轻人相似的神经表征(Zhang et al., 2023)。

根据前文所述, 与歌唱相关的音乐训练, 基本满足嗓音模仿的认知过程, 具体来说, 歌唱家或者乐器演奏家之所以可以产生或演奏动听的音乐, 其基础在于将感知到的目标音高, 反复地基于其产出相关器官, 进行再现, 这一过程需要感知运动相关脑区的参与。未来研究可以结合嗓音模仿任务, 来进一步探究音乐训练老年人的嗓音模仿能力, 旨在揭示嗓音模仿在言语老化中的作用。

5.3 嗓音模仿多阶段加工模型对模仿认知可塑性以及言语交流中的应用启示

如图 1 所示, 嗓音模仿具体包括嗓音感知, 嗓音感知映射产出, 产出指令以及发声器官运动四个阶段, 这些阶段对于嗓音模仿认知加工都是必要的, 不同阶段出现加工障碍都会影响说话人的嗓音模仿表现。根据前文所述, 嗓音模仿表现指标一方面通过计算说话人与目标说话人之间声学特征的相似性[17], 另一方面通过测量说话人与目标说话人发声器官运动表征的一致性[18]。在此基础上引申出的一个问题是, 如果对嗓音模仿中不同加工阶段进行训练, 那么特定加工阶段的训练效应可能会体现在嗓音模仿表现的提高, 那么这

[14] 这些音乐训练, 主要包含了歌唱, 乐器。

[15] 经过音乐训练和没有经过训练的老年人, 其年龄都在 65 岁左右。

[16] 没有经过训练的年轻人, 其年龄都在 23 岁左右。

[17] 说话人基线条件的声学参数与目标声学参数之间的欧式距离减去说话人模仿条件与目标声学参数之间的欧式距离(|Target-Baseline| − |Target-Imitation|), 如果模仿表现指数为负, 表明说话人与目标说话人嗓音不相似, 如果模仿表现指数为正, 表明说话人与目标说话人嗓音相似。

[18] 一致性是指说话人受到目标说话人发声器官运动表征与其实际发声器官运动表征一致或不一致的影响, 继而调节说话人言语产出延迟的时间。

种训练效应是否会迁移到别的加工阶段, 从而提高嗓音模仿表现。这种特定加工阶段的训练效应对其他加工阶段的迁移性, 也可以称为嗓音模仿的可塑性。未来研究可以通过训练嗓音辨别能力或发声控制能力, 使其有效提高说话人的嗓音模仿表现, 并且, 在此基础上进一步揭示感知映射发声运动的认知加工阶段是否会受到训练迁移效应的影响。对于这一问题的探讨, 有助于为唱音障碍群体提供更有效的干预训练策略, 从而减弱感知映射产出过程中出现的认知加工障碍。

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。

参考文献

蔡笑, 张清芳. (2020). 言语运动系统中前馈和反馈控制整合加工的作用机制. *心理科学进展*, *28*(4), 588−603.

Adank, P., Hagoort, P., & Bekkering, H. (2010). Imitation improves language comprehension. *Psychological Science*, *21*(12), 1903−1909.

Belyk, M., Brown, R., Beal, D. S., Roebroeck, A., McGettigan, C., Guldner, S., & Kotz, S. A. (2021). Human larynx motor cortices coordinate respiration for vocal-motor control. *NeuroImage*, *239*, 118326.

Belyk, M., & Brown, S. (2017). The origins of the vocal brain in humans. *Neuroscience & Biobehavioral Reviews*, *77*, 177−193.

Belyk, M., Eichert, N., & McGettigan, C. (2021). A dual larynx motor networks hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1840), 20200392.

Belyk, M., Johnson, J. F., & Kotz, S. A. (2018). Poor neuro-motor tuning of the human larynx: A comparison of sung and whistled pitch imitation. *Royal Society Open Science*, *5*(4), 171544.

Belyk, M., Lee, Y. S., & Brown, S. (2018). How does human motor cortex regulate vocal pitch in singers? *Royal Society Open Science*, *5*(8), 172208.

Belyk, M., & McGettigan, C. (2022). Real-time magnetic resonance imaging reveals distinct vocal tract configurations during spontaneous and volitional laughter. *Philosophical Transactions of the Royal Society B*, *377*(1863), 20210511.

Belyk, M., Pfordresher, P. Q., Liotti, M., & Brown, S. (2016). The neural basis of vocal pitch imitation in humans. *Journal of Cognitive Neuroscience*, *28*(4), 621−635.

Belyk, M., Schultz, B. G., Correia, J., Beal, D. S., & Kotz, S. A. (2019). Whistling shares a common tongue with speech: Bioacoustics from real-time MRI of the human vocal tract. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1911), 20191116.

Bernhold, Q. S., & Giles, H. (2020). Vocal accommodation and mimicry. *Journal of Nonverbal Behavior*, *44*(1), 41−62.

Bono, D., Belyk, M., Longo, M. R., & Dick, F. (2022). Beyond language: The unspoken sensory-motor representation of the tongue in non-primates, non-human and human primates. *Neuroscience & Biobehavioral Reviews*, *139*, 104730.

Brown, S., Yuan, Y., & Belyk, M. (2021). Evolution of the speech-ready brain: The voice/jaw connection in the human motor cortex. *Journal of Comparative Neurology*, *529*(5), 1018−1028.

Cartei, V., Oakhill, J., Garnham, A., Banerjee, R., & Reby, D. (2020). "This is what a mechanic sounds like": Children's vocal control reveals implicit occupational stereotypes. *Psychological Science*, *31*(8), 957−967.

Chen, T., Lammert, A. C., & Parrell, B. (2021). Modeling sensorimotor adaptation in speech through alterations to forward and inverse models. *Interspeech*, 3201−3205.

Chomsky, N., & Lightfoot, D. W. (2002). *Syntactic structures*. Walter de Gruyter.

Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics*, *90*, 101123.

Cracco, E., Bardi, L., Desmet, C., Genschow, O., Rigoni, D., de Coster, L., ... Brass, M. J. P. B. (2018). Automatic imitation: A meta-analysis. *Psychological Bulletin*, *144*(5), 453−500.

Drake, E., & Corley, M. (2015). Articulatory imaging implicates prediction during spoken language comprehension. *Memory & Cognition*, *43*(8), 1136−1147.

Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology*, *4*, 346.

Frühholz, S., & Schweinberger, S. R. (2021). Nonverbal auditory communication − Evidence for integrated neural systems for voice signal production and perception. *Progress in Neurobiology*, *199*, 101948.

Galantucci, B., Fowler, C. A., & Goldstein, L. (2009).

Perceptuomotor compatibility effects in speech. *Attention, Perception, & Psychophysics*, *71*(5), 1138−1149.

Gambi, C., van de Cavey, J., & Pickering, M. J. (2022). Representation of others' synchronous and asynchronous sentences interferes with sentence production. *Quarterly Journal of Experimental Psychology*, *76*(1), 180−195.

Gandolfi, G., Pickering, M. J., & Garrod, S. (2022). Mechanisms of alignment: Shared control, social cognition and metacognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1870), 20210362.

Garnier, M., Lamalle, L., & Sato, M. (2013). Neural correlates of phonetic convergence and speech imitation. *Frontiers in Psychology*, *4*, 600.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251−279.

Herbst, C. T. (2020). Electroglottography − An update. *Journal of Voice*, *34*(4), 503−526.

Heyes, C. (2001). Causes and consequences of imitation. *Trends in Cognitive Sciences*, *5*(6), 253−261.

Heyes, C. (2011). Automatic imitation. *Psychological Bulletin*, *137*(3), 463−483.

Heyes, C. (2021). Imitation. *Current Biology*, *31*(5), R228−R232.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131−138.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1-2), 67−99.

Janssen, N., Kessels, R. P. C., Mars, R. B., Llera, A., Beckmann, C. F., & Roelofs, A. (2023). Dissociating the functional roles of arcuate fasciculus subtracts in speech production. *Cerebral Cortex*, *33*(6), 2539−2547.

Kim, D., & Clayards, M. (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769−786.

Kinzler, K. D. (2021). Language as a social cue. *Annual Review of Psychology*, *72*(1), 241−264.

Kuhlen, A. K., & Abdel Rahman, R. (2023). Beyond speaking: Neurocognitive perspectives on language production in social interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1875), 20210483.

Lin, J.-F. L., Imada, T., Meltzoff, A. N., Hiraishi, H., Ikeda, T., Takahashi, T., ... Kuhl, P. K. (2023). Dual-MEG interbrain synchronization during turn-taking verbal interactions between mothers and children. *Cerebral Cortex*, *33*(7), 4116−4134.

Liu, Y., Zhao, Z., Xu, M., Yu, H., Zhu, Y., Zhang, J., ... Wu, J. (2023). Decoding and synthesizing tonal language speech from brain activity. *Science Advances*, *9*(23), eadh0478.

Mai, G., & Howell, P. (2023). The possible role of early-stage phase-locked neural activities in speech-in-noise perception in human adults across age and hearing loss. *Hearing Research*, *427*, 108647.

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is production: The missing link between language production and comprehension. *Scientific Reports*, *8*(1), 1079.

Mercado, E., Mantell, J. T., & Pfordresher, P. Q. (2014). Imitating sounds: A cognitive approach to understanding vocal imitation. *Comparative Cognition & Behavior Reviews*, *9*, 1−57.

Mousikou, P., Strycharczuk, P., Turk, A., & Scobbie, J. M. (2021). Coarticulation across morpheme boundaries: An ultrasound study of past-tense inflection in Scottish English. *Journal of Phonetics*, *88*, 101101.

Nguyen, N., & Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. *Journal of Phonetics*, *53*, 46−54.

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language*, *69*(3), 183−195.

Pardo, J. S., Pellegrino, E., Dellwo, V., & Möbius, B. (2022). Special issue: Vocal accommodation in speech communication. *Journal of Phonetics*, *95*, 101196.

Pardo, J. S., & Remez, R. E. (2021). On the relation between speech perception and speech production. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The Handbook of Speech Perception* (pp.632−655). Wiley Online Library. https://doi.org/10.1002/9781119184096.ch23

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637−659.

Paroni, A., Henrich Bernardoni, N., Savariaux, C., Lœvenbruck, H., Calabrese, P., Pellegrini, T., ... Gerber, S. (2021). Vocal drum sounds in human beatboxing: An acoustic and articulatory exploration using electromagnetic articulography. *The Journal of the Acoustical Society of America*, *149*(1), 191−206.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, *333*(6042), 595.

Peschke, C., Ziegler, W., Kappes, J., & Baumgaertner, A. (2009). Auditory−motor integration during fast repetition: The neuronal correlates of shadowing. *NeuroImage*, *47*(1), 392−402.

Pfordresher, P. Q., & Mantell, J. T. (2014). Singing with yourself: Evidence for an inverse modeling account of poor-pitch singing. *Cognitive Psychology*, *70*, 31−57.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological Bulletin*, *144*(10), 1002−1044.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic

psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169−190.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), 329−347.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: A window into the origins of human vocal control? *Trends in Cognitive Sciences*, *20*(4), 304−318.

Shmuelof, L., & Krakauer, J. W. (2011). Are we ready for a natural history of motor learning? *Neuron*, *72*(3), 469−476.

Stansbury, A. L., & Janik, V. M. (2019). Formant modification through vocal production learning in gray seals. *Current Biology*, *29*(13), 2244−2249.e4.

Stoeger, A. S., Mietchen, D., Oh, S., de Silva, S., Herbst, C. T., Kwon, S., & Fitch, W. T. (2012). An Asian elephant imitates human speech. *Current Biology*, *22*(22), 2144−2148.

Virhia, J., Kotz, S. A., & Adank, P. (2019). Emotional state dependence facilitates automatic imitation of visual speech. *Quarterly Journal of Experimental Psychology*, *72*(12), 2833−2847.

Waters, S., Kanber, E., Lavan, N., Belyk, M., Carey, D., Cartei, V., ... McGettigan, C. (2021). Singers show enhanced performance and neural representation of vocal imitation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *376*(1840), 20200399.

Wilt, H., Wu, Y., Trotter, A., & Adank, P. (2023). Automatic imitation of human and computer-generated vocal stimuli. *Psychonomic Bulletin & Review*, *30*(3), 1093−1102.

Wu, Y., Evans, B. G., & Adank, P. (2019). Sensorimotor training modulates automatic imitation of visual speech. *Psychonomic Bulletin & Review*, *26*(5), 1711−1718.

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "autistic" traits. *PLOS ONE*, *8*(9), e74746.

Zellou, G., Cohn, M., & Kline, T. (2021). The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience*, *36*(10), 1298−1312.

Zhang, J., Liu, D.-Q., Qian, S., Qu, X., Zhang, P., Ding, N., & Zang, Y.-F. (2022). The neural correlates of amplitude of low-frequency fluctuation: A multimodal resting-state MEG and fMRI−EEG study. *Cerebral Cortex*, *33*(4), 1119−1129.

Zhang, L., Wang, X., Alain, C., & Du, Y. (2023). Successful aging of musicians: Preservation of sensorimotor regions aids audiovisual speech-in-noise perception. *Science Advances*, *9*(17), eadg7056.

Multi-stage model of neurocognitive processing for vocal imitation

HU Yanbing[1], JIANG Xiaoming[1,2]

([1] *Institute of Linguistics, Shanghai International Studies University, Shanghai* 201620, *China*)

([2] *Key Laboratory of Language Science and Multilingual Artificial Intelligence, Shanghai International Studies University, Shanghai* 201620, *China*)

Abstract: Vocal imitation is a fundamental cognitive process in verbal communication, where the speaker maps the voice signal of another (the target speaker) onto their own vocal organ motion representation. This process aims to replicate the vocal organ motion and reproduce the target speaker's voice. Neuroimaging studies reveal that this cognitive processing involves a neural network extending from the superior temporal gyrus to the left inferior frontal gyrus, and finally to the vocalization-associated primary motor cortex, with the basal ganglia coordinating this network. Variations in voice discrimination ability, the capacity for mapping voice signals to vocal motor representations, and the control of vocal organ muscles significantly influence this cognitive process. Future research should focus on integrating vocal imitation studies with vocal disorders and intracranial electrode technology. This integration aims to uncover the causal mechanisms between brain function and behavior underlying vocal imitation and apply these insights to lifelong speech development, cognitive plasticity, and advancing the field of speech anticipation.

Keywords: vocal imitation, vocal reenactment, vocal motor movements copy, imitation neural network, individual differences