语音智能体：从听觉智能到听觉能动性

Kele Xu，资深会员,电气与电子工程师协会高级会员，Qisheng Xu，Xinwang Liu，电气与电子工程师协会高级会员，Huaimin Wan g
电气与电子工程师
协会

arXiv:2611.29465v1 [eess.AS] 22 Nov 2026

摘要——**语音系统日益融合流式语音、持久状态、推理、工具使用和具身控制。当这些功能共用同一运行时时，语音输入可能改变数字或物理状态；仅凭识别或响应质量已无法全面刻画系统。本文探讨语音证据产生实际影响所经由的接口。**听觉智能**指感知、理解、协调并生成语音交互所需的能力；**听觉能动性**指从这种交互到有状态决策及可观察后果的因果路径。**承诺边界**标示可修订的听觉证据转变为持久状态、计划或可执行请求的位置。随后，**能动性闭环**沿着经授权的行动，依次涵盖观测、验证、恢复或补偿，以及结果的交流。来自全双工交互、推理、记忆、工具、具身性和人工控制的证据表明，系统层面存在一种持续失配：集成的推进速度快于对集成系统实际行为的端到端观测。因此，设计问题在于保留从语音证据到后果的追踪链，同时确保承诺、执行和验证均可追溯。**

索引词—**语音智能体、口语对话系统、语音基础模型、听觉能动性、具身智能体**

I. 引言

近年来，语音系统的能力已远远超出语音识别和响应生成的范畴。有些系统能够边听边说，另一些则能在多个会话中保留信息、调用外部工具，或将口头指令连接到具身控制器 [1]–[4]。这方面的许多进展都是通过单项能力来评估的，而对这些能力如何协同运作关注较少。当口头请求尚未说完，或正在被更正时，系统却已开始准备执行动作，这一难题便显现出来。即使识别准确、响应流畅，也无法证明请求是否得到了正确理解、动作是否获得授权，或最终的外部状态是否符合用户的意图。

这一区分促使我们采用两个相关概念。我们用 听觉智能来描述语音交互[5]中涉及的感知、语言和对话能力。相比之下，听觉能动性关注这些能力与有状态决策及可观察后果的结合程度。现有系统从不同方向探索这一问题。Moshi、SALM-Duplex 和 Freeze-Omni 研究语音重叠与交互时序 [1], [6], [7]。AURA 和 AnovaX 引入了规划或执行接口 [3], [8], ，而具身系统则将语音指令与导航和操作相连 [4], [9]。这些进展引出了一个共同的系统问题：智能体应如何将尚未确定的语音输入转化为授权承诺，又该如何判断预期动作是否确已执行?

A. 为什么语音智能体已经成为一个系统问题

语音原生基础模型让声学处理与语言建模更加紧密地结合起来。AudioPaLM 展示了如何在共享模型中联合建模语音与语言 [10],，而 SALMONN- omni 和 MinMo 则推动了流式处理与统一语音交互的发展 [11], [12]。与此同时，语义端点检测和增量理解使系统能够在说话者尚未说完时，解读话语中有意义的部分 [13], [14]。Moshi、SALM-Duplex 和 Freeze- Omni 进一步支持重叠语音输入与输出，突破了听与说轮流进行的传统模式 [1], [6], [7]。因此，说话者仍在组织或修正请求时，语音系统便可以开始理解并回应。

其影响不止于对话时序。增量语义信息可以在完整转录文本生成之前启动推理 [15], [16]。记忆机制可以在不同会话之间保留用户偏好和交互历史，而身份感知设计则力求确保这些信息与相应用户关联 [2], [17], [18]。语音到函数模型和同步动作流为口头请求与结构化

命令 [19], [20], 多模态与具身系统将语音连接到导航、操作和情境化协助 [4], [9], [21], [22]。这些能力使解释、推理、记忆和动作之间的联系日益紧密。对话语早期形成的解释，如今可能在说话者的意图完全明确之前，就影响决策或下游行为。

管理此类依赖并非全新之事。经典口语对话系统早已将对话状态、策略、服务访问和语音合成整合在一起，通常通过具有明确接口和职责的模块来实现 [23]–[25]。基础模型架构使这些功能能够更紧密地协同，且感知、对话管理和规划之间往往不再有同样明确的边界。这种整合可以让语音交互更加流畅，但也使中间决策的责任归属更难厘清。系统仍必须确定，何时临时解释已充分确立，足以支持经授权的行动

Kele Xu 任职于中国长沙 410073 国防科技大学计算机学院，以及中国长沙 410073 复杂与关键软件环境国家重点实验室（电子邮箱：kelele.xu@gmail.com；ORCID：0000-0002-7136-1898）。

由哪个组件负责该决策，以及应如何处理后续修订。

B. 听觉能动性、承诺与闭环

交互智能与智能体智能是两个相互独立的分析维度。交互智能涉及参与口语交流的能力——感知、来源与受话者归因、增量理解、轮次协调、语音生成以及修复。智能体智能涉及超越当前话语而持续存在的目标导向状态——推理、规划、记忆、授权、动作、观测、验证与恢复。两个维度都不是成熟度等级。系统问题出现在二者交汇之处，即对话证据开始承载后果之时。

我们将这一节点称为 承诺边界。在承诺之前，关于词语、意图、受话者或目标的假设都可以保持为暂定状态。在承诺之后，某种解释已被写入持久状态、用于实例化计划，或被转换为符合执行条件的请求。在流式系统中，声学端点与承诺决策不一定重合。前者询问的是语音是否似乎已经停止；后者询问的是，是否存在足够的证据，使下游状态或动作可以依赖当前解释。将这两个事件分开，有助于更容易识别过早动作，并且原则上也更容易撤回这些动作。

第二个概念是<styled id='1'>能动性闭环。意图中的动作获得授权并被执行；随后对由此产生的外部状态进行观测，并与预期状态进行核对；不匹配会触发恢复或补偿；结果则会反馈给用户。这一定义将几个经常被混为一谈的事件区分开来。工具选择并不等同于工具执行。执行也不能证明预期状态已经发生改变。对状态的观测还不等于验证用户的目标已经达成。在交流以及多模态任务导向协作中，共同理解的建立与修复长期以来一直被作为共同基础加以研究 [26], [27]。

记忆、授权和恢复自然地融入这一论述。持久记忆承载着身份、来源和时间范围。授权决定哪些原本有效的计划可以进入执行阶段。恢复决定在状态转换失败或部分完成后将发生什么。后续章节将回到四个具体问题：承诺了什么，谁或什么拥有决策权，什么证据确立了后果，以及当预期状态与观测到的状态不一致时发生了什么。

C. 范围与贡献

此处的 语音智能体 指一种系统：它接收语音或更丰富的听觉输入，维护交互状态或任务状态，选择响应或动作，并生成语音或其他可观察到的后果。本文的组织单元并非产品类别或模型系列，而是听觉证据转化为持久状态、决策、动作或结果所经过的接口。经典对话系统、语音语言模型、双工交互、记忆、工具使用、具身控制和以人为中心的评估，因此会在有助于阐明上述某种接口时加以讨论。

分析沿着从口头请求到其后果的路径展开。承诺边界和能动性闭环为这一路径提供了两个参照点。架构的比较依据是状态、决策、执行和验证由谁负责；评估结果则根据其实际观察的终点来解读。这样既能让语音中的不确定性、权限、持久状态和外部后果清晰可见，也无需将它们强行纳入同一成熟度量表。

讨论将从架构（第II节）和交互智能（第III节）延伸至智能体智能（第IV节）、以人为中心的控制（第V节）、闭环评估（第VI节），以及由这些接口引出的研究议程（第VII节）。

II. 语音智能体的演进与架构

语音智能体能否修正早期的解释，或从失败的动作中恢复，部分取决于其架构。模块化系统会在接口处显露不确定性和责任归属；集成模型则可能让这些边界更难以界定。语音生成、流式理解、轮次控制、检索、推理和安全性的发展在一定程度上各自独立，如今正逐步整合到共同的运行时环境中。随着这些功能开始共享状态，决策、授权和恢复的安排便成为一个核心设计关注点。

A. 从口语对话系统到基础模型

识别加策略。早期口语对话系统将识别、对话状态、策略选择、后端访问和合成组织为清晰可辨的阶段。概率式对话管理将交互视为部分可观测性下的决策过程，通过信念状态表示不确定性，并依据显式奖励优化策略 [28]。Plato 展示了模块化方法，将对话状态、策略和后端服务作为可替换组件公开 [23]。状态跟踪也开始更加关注语音信号中的不确定性。词混淆网络将备选识别假设纳入状态估计 [24], ，而时间表达网络和在线强化学习则探索了信念跟踪与策略的联合优化 [25], [29], [30]。这些系统奠定了状态、策略和奖励等术语基础，这些术语至今仍用于语音智能体。它们的显式接口也使中间决策相对容易检查，不过评估通常仍局限于有限领域和特定任务目标。

神经建模与原生语音建模。神经建模改变了这些接口所能传递的内容。语音合成技术的进步，使生成的语音更适合用于对话 [31],；与此同时，语音到语音翻译也越来越多地避免使用纯文本中间表示。Translatotron 2 解决了翻译过程中的语音保留问题 [32]；后续方法则采用了学习得到的离散单元 [33],、两遍解码 [34], 和单语训练数据 [35],，之后 SeamlessM4T 又将其扩展至多语言环境 [36]。

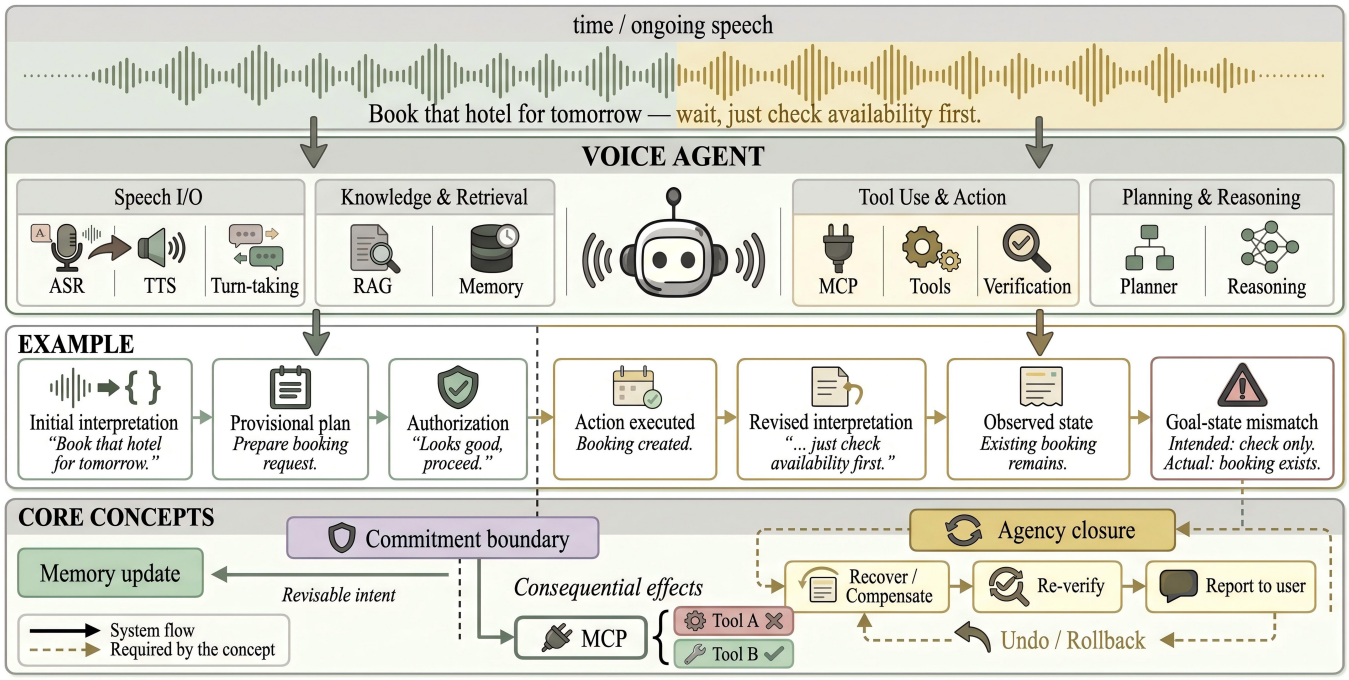

图1. 从听觉智能到听觉能动性。该图说明，部分语音输入可能触发过早的动作，与用户随后提出的更正相冲突。图中突出显示了暂定解释与下游状态或动作之间的承诺边界，并介绍了能动性闭环：验证后果、处理不匹配情况，并向用户报告结果。实线箭头描绘了示例性失败场景，虚线箭头则表示所需的控制与恢复路径。

AudioPaLM 将语音输入与生成纳入统一的语言模型形式 [10]。SpeechGPT 将指令遵循扩展到离散语音表示 [37],; SpiRit-LM 则交错使用书面与口语词元，并在其富有表现力的变体中加入音高和风格单元 [38]。双通道生成式对话模型还表明，无须单独指定轮流发言模块，也能对交互式语音进行建模 [39]。这些进展拓展了语音系统的声学与语言能力。但仅凭这些进展，并不能解决系统应如何识别受话者、维护用户专属状态，或在说话者改变主意后撤销动作等问题。

流式与时间控制。另一条相关研究路线探讨了语音以增量方式到达这一事实。早期的增量式对话架构将部分解释表示为关联单元，这些单元可以被添加、修订或确认，从而使纠正能够沿依赖关系传播至各个模块 [40]。增量生成将类似原则应用于输出端：系统可以在规划完整响应之前就开始语音表达 [41]。随后，流式口语理解能够根据不完整的音频生成意图和槽位预测 [13],，响应时机模型则研究系统何时应开始语音表达，以及应该说什么 [42], [43]。将 TurnGPT 与语音活动预测相结合的研究，把学习得到的时序预测应用于对话机器人，涵盖准备、轮流发言和打断处理 [44]。端点检测也不再局限于静音检测。Phoenix-VAD 利用语义信息判断话语是否结束 [14]；自适应检测则处理由残余播放导致的误判打断 [45]；融合模型能够区分停顿、反馈语、打断，以及续接话语 [46]。设备指向性提出了另一个问题：这段语音一开始是否就是对助手说的 [47]。总的来说，这些机制让系统能够更早作出响应，但也使修订变得更加重要。迟到的纠正可能会在推理已经开始或操作已经准备就绪之后才到来。

语音语言模型与双工对话。近期的语音语言模型进一步拉近了声学表示、语言建模和对话控制之间的联系 [11], [12]。双工架构更进一步，使听觉理解与语音表达能够重叠进行。Moshi 对用户和系统的编解码器词元流进行并行建模，同时处理内部文本流，据报告，其理论延迟约为 160 毫秒 [1]。SALM-Duplex 学习何时听、何时说 [6], ，而 Freeze-Omni 则为冻结语言模型增添了双工交互能力 [7]。GLM-4-Voice 探索端到端口语对话 [48], ，相关研究则将语音语言建模扩展至更长的上下文 [49]。其他设计保留了彼此独立的组件：LLaMA-Omni 2 将语言模型骨干网络与语音编码器和流式自回归解码器相结合 [50]；SyncLLM 引入时间信息，使生成与现实时间同步 [51]；OmniFlatten 则采用展平的多模态序列和分阶段训练，以处理重叠发言、反馈语和打断 [52]。双工控制也可以保留为可插拔模块 [53], ，而结构化动作模型则引入独立的流来进行规划和工具调用 [20]。随着这些变化，评估范围也不断扩展，从轮流发言和重叠处理 [54]，到自动化多轮测试 [55], 、结合模拟工具 API 的含不流畅现象的人类录音 [56],、富含打断的对话生成 [57]，以及打断后的恢复 [58]。由此产生的基准测试

揭示了对话时序的不同方面；但单独来看，没有任何一项能确定某个外部操作是否恰当或已完成。

推理、工具、记忆与共同基础。随着语音接口变得更加灵敏，研究人员开始将其与更长的任务序列相连接。在更广泛的语言智能体文献中提出的 ReAct，将推理、动作和观测交织在一起，为面向工具的控制循环提供了重要先例 [59]。面向语音的研究通过语言模型控制器协调专用音频模型 [60],，从增量语义事件启动推理 [16], ，并将思维链训练应用于级联式和流式双工对话 [15], [61]。语音转函数基准测试评估将含噪或不流畅的请求转换为结构化调用的能力 [19], [62]。智能体框架明确呈现编排和类型化执行 [3], [8], ，而基于实际状态的基准测试则将数据库最终状态与指定目标进行比较 [63]。面向领域的系统展示了表单填写、对话式搜索和企业支持的实际需求 [64]–[66]。在语音之外， τ-bench 使用领域策略、API 工具、模拟用户和重复试验来评估工具使用型智能体，并依据带注释的最终数据库状态衡量任务完成情况 [67]。此类评估使工具使用的后果得以显现，尽管模拟调用与外部服务中经验证的更改仍是不同形式的证据。

近期的多模态发布版本展示了这些功能如何在模型层面逐渐融合。小米的 MiMo 文档介绍了同一模型系列中的原生音频理解和智能体执行 [68]。Qwen3.5-Omni 技术报告同样提出了一个共享基础，支持音视频理解、流式语音生成，以及原生工具接口，包括 WebSearch 和 FunctionCall [69]。这些描述的是模型和平台的能力，并非外部操作已执行并经过验证的独立证据。不过，这也使时序、记忆、工具使用和结果检查的所有权问题变得更加迫切。

持久性、具身性与治理。另一系列进展关注的是，哪些内容会在一轮对话结束后留存，以及对话之外会发生什么。检索增强系统将外部知识引入对话 [70]–[72]；记忆机制则提取并再利用跨会话的偏好 [2], [17]；长时对话音频也揭示了可测量的性能退化 [73]。尽管 LongMemEval 基于文本，却能测试信息提取、多会话推理与时间推理、知识更新及拒答，为评估持久状态提供了有用的区分维度 [74]。持久性也带来了安全与隐私方面的顾虑。语音特定攻击与防御措施 [75], [76], 、交互隐私基准测试 [77], ，以及针对使用工具的语音助手的治理框架 [78] ，分别考察了这一问题的不同方面。物理接地又增加了一项约束。SayCan 源自语言引导的机器人技术，它将语言模型知识与技能价值函数相结合，从而选择对机器人及其所处环境而言可行的动作 [79]。涉及驾驶模拟器、社交机器人和临床评估的研究表明，当语音决策会影响情境化任务时，这类接口为何至关重要 [9], [80]–[81]。

这些历史脉络对于当今系统仍然具有现实意义。翻译质量、富有表现力的合成、低资源覆盖和安全性，仍然塑造着推理与动作所依赖的交互。不过，它们的评估指标描述的是不同的终点。工具选择 F 值、轮次状态分类准确率，以及真实交互条件下的任务完成情况，不能被视为衡量能动性的可互换指标。因此，比较架构时需要考察每个系统能让哪些内容变得可观察，从语音行为到已执行的行动及其后果。

B. 架构类别

主要的架构差异在于信息与控制传递所经由的接口。级联架构通常交换文本和结构化对话状态；语音原生系统可以保留学习到的声学表示；双工设计协调并发的音频与控制信号；而面向智能体的系统则增加了调用、观测和状态更新。这些是宽泛的设计类别，而非互斥的分类。表 I 对比了它们对状态和决策归属的处理，以及使这些边界具有时序敏感性的交互机制。

级联架构与部分集成流水线。 传统的级联架构将声学前端、识别、语言理解、对话管理、后端服务和合成分隔开来。只要各阶段可观测，中间输出就能让错误更容易定位。其代价是延迟累积、错误传播，以及将语音简化为文本时韵律或重叠信息的丢失。部分集成可以在不消除所有边界的情况下解决其中一些问题。对话合成根据对话历史 [31]调整生成；流式理解从部分输入 [13]中提取意图和槽位；双工封装器为顺序运行的后端增加并行的听觉理解与语音表达 [53]；混合控制器则将快速交互处理与较慢的推理分开 [16]。

基于令牌的架构与语音语言架构。许多原生语音架构以学习得到的离散单元、编解码器令牌或直接音频表示取代纯文本接口。离散单元方法和双通道方法使用学习得到的语音单元进行生成 [33]–[35]。AudioPaLM 在共享模型中处理并生成语音 [10]；SALMONN-omni 将流式编码与解码结合起来，不存在编解码器瓶颈 [11]；统一多模态模型则将语音处理扩展到各类交互任务 [12]。保留声学信息能够留存纯转录文本会舍弃的线索，也可能降低接口开销。代价是内部决策过程不那么透明，尤其是在规划或工具使用被整合进同一模型时。像 Moshi 这样采用中间文本流的架构，能够提高一定的可见性，同时不必明确呈现整个行动路径 [1]。

端到端架构与双工架构。双工系统将听觉理解和语音表达视为重叠活动，而非严格交替的轮次。双通道生成式对话模型确立了一种方案 [39]；Moshi 对并行语音流进行建模 [1]；SALM-Duplex 学习听或说的行为 [6]；Freeze-Omni 则在保留冻结语言模型的同时，学习双工语音接口

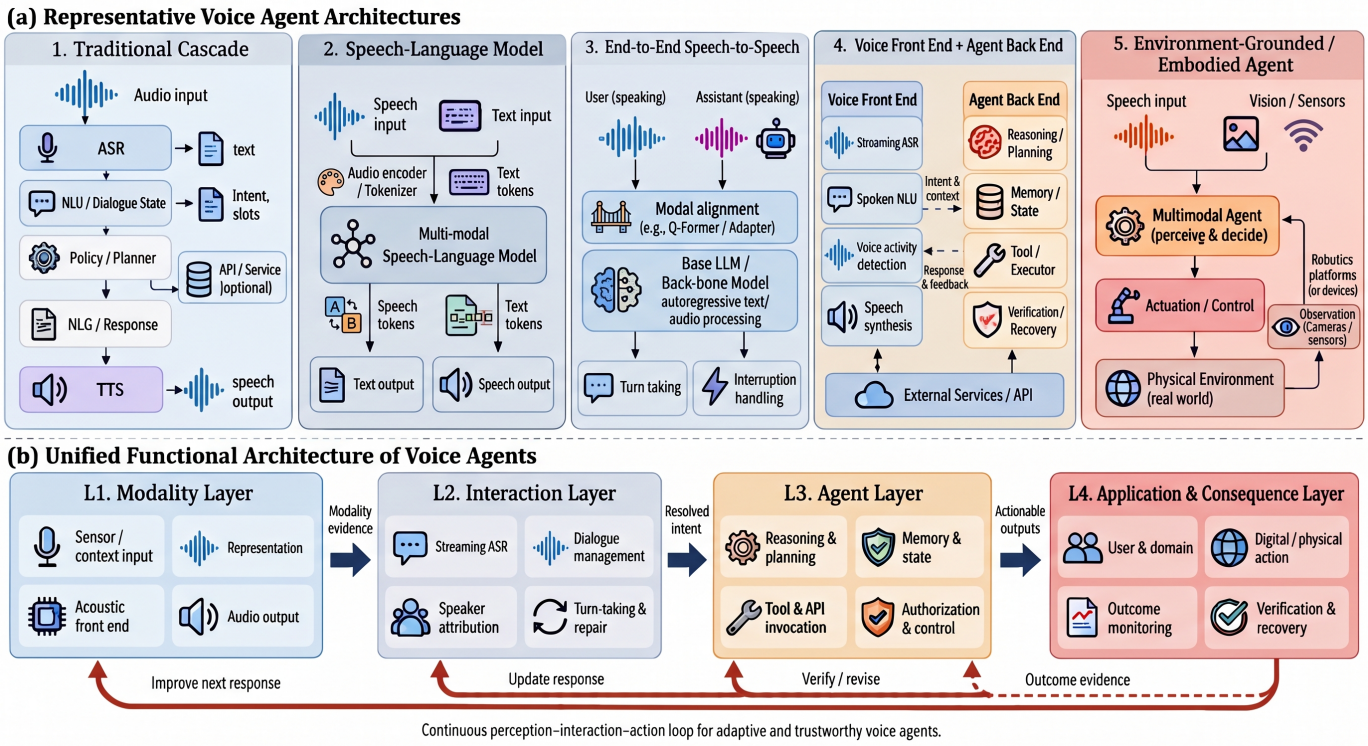

图2. 语音智能体的代表性架构范式与统一功能组织。(a) 五种代表性架构范式，涵盖从传统级联架构和语音语言模型，到智能体集成系统与环境落地系统。(b) 统一的四层功能架构，涵盖模态处理、交互管理、智能体推理与控制以及应用层后果，并通过结果驱动反馈支持持续适应与验证。

[7]。轮次级交错在集成建模与交互控制之间提供了另一种平衡 [87]。在这些架构中，轮流发言可以成为生成的一部分，而不再是在生成回应之前单独作出的决策。这有望促成流畅对话，但也可能使人更难将流畅却不正确的回应追溯到某个具体的中间决策。

采用智能体后端的语音前端。实现面向任务的语音交互的一种常见方式，是将负责端点检测、并发和轮次控制的前端与负责推理和执行的后端配对。AnovaX 描述了一种由语音前端、可选规划器、类型化执行器、安全层和生命周期事件总线组成的架构 [8]；AURA 同样将理解、推理和自动化工具使用分开处理 [3]。面向特定领域的系统也采用了类似的安排，用于表单填写、搜索和支持 [64]–[66], [88]。明确的动作接口有助于定位故障，并限制可尝试的操作。这些限制也会影响报告的可靠性：执行器的功能越窄，发生故障的机会就越少。更根本的是，即使工具调用格式正确，也可能未经授权、被服务拒绝，或虽已执行却未能产生预期状态。

落地与具身架构。物理环境让语音与动作之间的交互尤为显著。DOROTHIE 将对话动作与导航决策交织在一起 [9]；其他机器人系统则利用全模态上下文实现主动操作 [21]，或支持服务机器人上的低延迟多模态交互 [89]。将注意力分配给语音与动作专家的系统，也揭示了并发带来的实际代价：边说话边行动会降低长时域任务性能 [4]。共同基础要求系统根据场景解析指代、在恰当时机进行修复，并确保控制能力始终可供人工干预。若干有代表性的评估仍依赖仿真以及脚本化或生成式交互，因此限制了它们对现实世界运行情况的验证力度 [9], [22]。

C. 跨栈智能与控制分布

不同架构类别在听觉理解、决策、记忆和行动的分配方式上存在差异。图 2(b) 将反复出现的功能组织为模态层、交互层、智能体层以及应用与后果层。表 I 展示了某一给定功能如何由单个模型承担，或分配给交互前端、推理器、控制器、执行器和记忆存储。这样的配置会影响系统行为以及可用于评估系统的证据。

常见的架构大致有三种。在集中式设计中，一个模型负责音频输入与输出，许多感知和对话决策则在内部完成 [1], [12]。显式交接环节较少，可以简化交互，但观察者可能无法区分某项决策是缺失了，还是做错了。分层式系统将快速的交互前端与较慢的推理或动作后端配对 [16], [53]。只要接口能够忠实传递纠正信息和目标，这种架构就能让动作接口保持易于访问，并有助于分配延迟。在分布式系统中，感知、交互控制、推理、记忆和执行由更多独立负责组件承担 [3], [8]。组件可以替换，故障也更容易定位，

表 I 架构与交互设计概况。上半部分比较状态、决策和后果由谁掌控；下半部分总结使语音交互能够实时进行并支持修订的机制。所报告的数字保留其原始实验范围。

模式	状态／机制	决策或证据边界	系统影响	代表性证据
A. 架构范式				
级联	文本和明确的符号状态 在模块之间交换	对话管理器和服务 层；应用程序观察结果	可见接口，但延迟和信息损失 延迟和信息损失	[23], [24]
语音语言 模型	音频、文本或潜在语音 在共享模型中表示	共享模型；工具和验证 在存在时仍处于外部	连贯的跨模态推理，以及 不那么明显的所有权边界	[10]–[12]
原生全双工	并发声学与语义 具有生成时 控制	联合生成器；具有实际影响的 行动通常仍处于下游	低延迟协调，但 交互与动作可能变得 纠缠在一起	[1], [6], [7]
分层语音智能体	流式交互前端 加上规划器和类型化执行器	规划器提出；执行器 作出承诺；应用程序或验证器 观察	明确承诺和 后果边界	[3], [8]
工具增强型智能体	语音或文本与 结构化调用和观测相结合	编排器加外部服务	能力迅速扩展； 授权和部分故障 成为首要关注事项	[19], [20], [56]
具身智能体	语音结合视觉， 场景状态和控制信号	多模态策略或规划器加上 控制器；传感器呈现世界 状态	后果是可观察的，但 可逆性和接管则更难实现	[4], [9], [21]
B. 实时全双工交互机制				
原生双工 生成	并行编解码流或学习到的 听说控制	并发聆听与生成	原生低延迟对话；交互 证据本身并不能确立 行动闭环	Moshi；SALM-Duplex； Freeze-Omni [1], [6], [7]
统一的多模态 双工	自回归或共享式 多模态语音建模	在共同框架内的并发 模型状态	将双工对话与 更广泛的语音任务相结合	BayLing-Duplex；MinMo [12], [82]
模块化双工 控制	叠加在一个 现有语音/语言后端之上的控制器 end	轮次策略仍与 后端相分离	全模型重新训练的可插拔替代方案 重新训练	FlexDuo；DuplexCascade [53], [83]
同步语音 + 动作	语音、规划、轮次标签和 结构化调用共用一个时钟	合法的调用时序和参数 匹配	体现了延迟–行动权衡； 执行与授权仍然 相互独立	DuplexSLA：85.56%，耗时 0.648 秒；而耗时 2.77 秒时为 91.33% 用于级联 [20]
Duplex基准测试 使用工具	重叠、言语不流畅、自我- 纠正和模拟API	交互和有界工具调用 追踪链	将时序与任务证据相连 而不声称产生了实际副作用	全双工基准测试 v2/v3： F_1 0.876；pass@1 0.600 [55], [56]
打断与 恢复	中断场景加上 恢复评分标准	中断后行为会单独评分 ，与常规任务 成功	使恢复成为独立的 终点	FD-Bench；IHBench [57], [58]
端点检测 / 预期	语义补全、插话， 多状态决策，以及早期 预测	临时解释与 已确定的解释的时序	使承诺边界 明确	[14], [46], [84]–[86]

但每增加一项契约，都可能丢失有关身份、不确定性或授权的信息。

这些选择正日益延伸至设备本身。高通的2026年 Snapdragon8 EliteGen 6 平台介绍了增强型感知中枢、端侧上下文处理，以及用于智能体工作负载的 AI引擎 资源。Extreme平台还宣传了用于语音与环境音分离的 AI Voice Bubble，以及对端侧 30B+ 混合专家模型 [90], [91] 的支持。二手报道援引高通的说法称，感知中枢支持小型模型运行，并且该平台能够运行完整的语音输入/语音输出智能体 [92]。这些仍是厂商报告的能力，而非对听觉能动性的独立演示。不过，它们确实将延迟、隐私、个性化和控制权归属等问题带到了设备与云端边界之间。

集中式系统可能会减少接口开销，而分层系统和分布式系统则提供更多可观测授权、执行和验证的环节。当做出决策的组件同时报告其成功时，这一区别就很重要：该报告并非对结果的独立观测。基于事实的评估和明确的执行接口说明了为什么可观测边界很重要 [8], [63]。有意义的设计问题是，针对特定应用程序，哪些边界需要保持可检查，而不是系统被称为集成式还是模块化。

D. 架构权衡

没有哪一种架构系列适用于所有场景。相关权衡取决于一项任务能够容忍的延迟、所需的声学信息，以及当某个动作会产生后果时，必须保留多少控制权和证据以供访问。

延迟。级联架构会在各个阶段产生延迟，不过部分处理可以并行进行；集成模型消除了交接环节，但生成过程仍然需要时间。只有明确延迟的起止点，报告的延迟才有意义。模型分块延迟、首音频延迟、端点到响应延迟和麦克风到动作延迟衡量的是不同时间区间，不应直接比较 [20], [56]。

模块化与可检查性。模块化设计可以将状态和策略作为独立且可检查的对象呈现 [23]。除非经过专门配置以便进行检测，否则集中式模型很少提供此类边界。ESPnet 等共享研究工具包支持可复现的组件级语音实验 [93], ，但要追踪智能体所做的重大决策，所需的可见性必须超越语音模型本身。

声学信息保留。文本接口会丢弃韵律、重叠以及其他非词汇线索，除非这些信息被单独编码。直接的语音到语音系统可以跨接口保留习得的语音表征 [33]；无编解码器的语音语言模型则通过另一种设计，让音频贯穿输入和输出路径 [11]。无论选择哪种方式，都不能保证下游环节会正确使用这些信息。

推理能力与记忆集成。将对话上下文置于靠近推理模型的位置，可以简化对不完整或不断变化的信息的使用 [15]。持久记忆则带来不同的要求：存储的偏好和历史记录可能需要独立于使用它们的模型进行检查、更正或删除 [2], [17]。便于进行推理的位置，不一定适合存放长期用户状态。

工具集成与部署成本。外部服务扩展了语音智能体的能力，但也增加了必须与模型计算一并管理的依赖项。语音到功能和同步动作系统提供了生成结构化调用的接口 [19], [20]。在已部署的系统中，延迟、可用性、权限和服务配置可能与调用本身的准确率同样重要。

故障隔离、安全性与可控性。分布式架构可以更轻松地隔离故障，并允许禁用或替换不安全组件 [78]。更高程度的集成会将更多责任交由模型承担，也可能使受控降级变得复杂。对于具有实际影响的行动，即使感知、推理和响应生成共用同一个模型，明确的授权边界仍然很有价值。

三、交互智能：从流式语音到全双工对话

上一节讨论的架构选择，在实时对话中尤为关键。语音智能体必须在语音到达时进行解读，判断是谁在说话，以及这段话是否是说给它听的，并回应可能在处理开始后才出现的打断或纠正。我们用 *交互智能* 来描述支持这一过程的感知与对话能力：感知、归因、增量理解、轮次预测、协调和修复。这些功能不必按固定顺序运行。它们的输出会随时间变化，某个阶段的修订可能会影响系统其他部分正在进行的推理、记忆或动作准备。

图 3 概述了这一组织方式。流式语音证据通过感知、归因和增量理解，持续融入可修订的交互状态。实时对话控制功能——包括端点检测、轮流发言、反馈、打断处理和生成——基于不断演变的状态运行，而候选响应或计划可以在承诺之前预先准备。后续的口语证据仍可能触发修复和重新规划。承诺边界标示出这样一个节点：越过该节点后，下游组件可能开始依赖当前解释，因此交互层打断和动作层取消属于不同的控制决策。

A. 感知：前端与识别

在近距离拾音、单说话人条件下，声学前端对于系统其他部分而言可能几乎不可见。当智能体自身的播放声、室内混响或其他说话人进入麦克风信号时，声学前端的作用便会显著凸显。回声消除和个性化增强会削弱播放声及干扰声学组件 [94], [95], ，而语音活动检测则判断是否存在语音。说话人特定活动检测还能进一步帮助区分有意打断与残余播放或背景对话 [45]。在识别开始之前，这一阶段的错误就可能改变交互进程。

识别仍然是上游不确定性的另一个重要来源。Conformer 为语音识别引入了卷积增强型 Transformer 架构 [96]，而 Whisper 则展现了大规模弱监督的价值 [97]。关于口音、方言、语码转换和低资源语音的研究，涵盖联合说话人分离与识别流水线、多语言评估、口音鲁棒检测，以及针对特定语言变体开发的系统 [98]–[102]。联邦式方法和端侧方法也致力于在无需将所有音频集中处理的情况下提升识别能力 [103]，[104]。然而，报告的错误率在很大程度上取决于测试集的领域、口音分布和噪声条件。更重要的是，对于智能体而言，识别错误造成的后果各不相同：即使话语的其余部分都转写正确，一个专有名词、数字或槽位值出错，也可能改变预期操作。这种敏感性有助于解释为何实体感知方法和检索增强方法持续受到关注 [70], [71]。

B. 归属判断：谁说了什么，以及对谁说

识别词语只是理解口头请求的一部分。在共享声学环境中，系统还必须确定该话语是否是对它说的，以及由哪位说话者发出。设备指向性解决的是前一个问题，有助于防止将人与人之间的对话误认为命令或打断 [47]。说话人分离和多说话人分离能够提供有关说话人身份以及信号时间归属的证据；高效的设备端说话人分离表明，此类处理可以在本地资源受限的条件下完成 [104]。多方对话中的受话人识别，以及针对交互隐私的基准测试，使这些区分成为明确的评估目标 [77], [105]。

Pyannote.audio 等工具支持将语音片段按时间分配给说话人 [106], ，而 x-vector 嵌入则为说话人识别提供表示 [107]。当系统能够执行动作时，这些技术的重要性也随之改变。在问答交互中，归因错误可能导致答非所问；在智能体中，则可能导致系统代表错误的人执行请求。即使说话人身份识别正确，该说话人也可能没有权限发起某项操作。因此，相关文献区分了说话人身份、验证和身份识别 [108]；这些概念本身都不能证明已获授权。我们将在第V节中重新讨论这一界限。

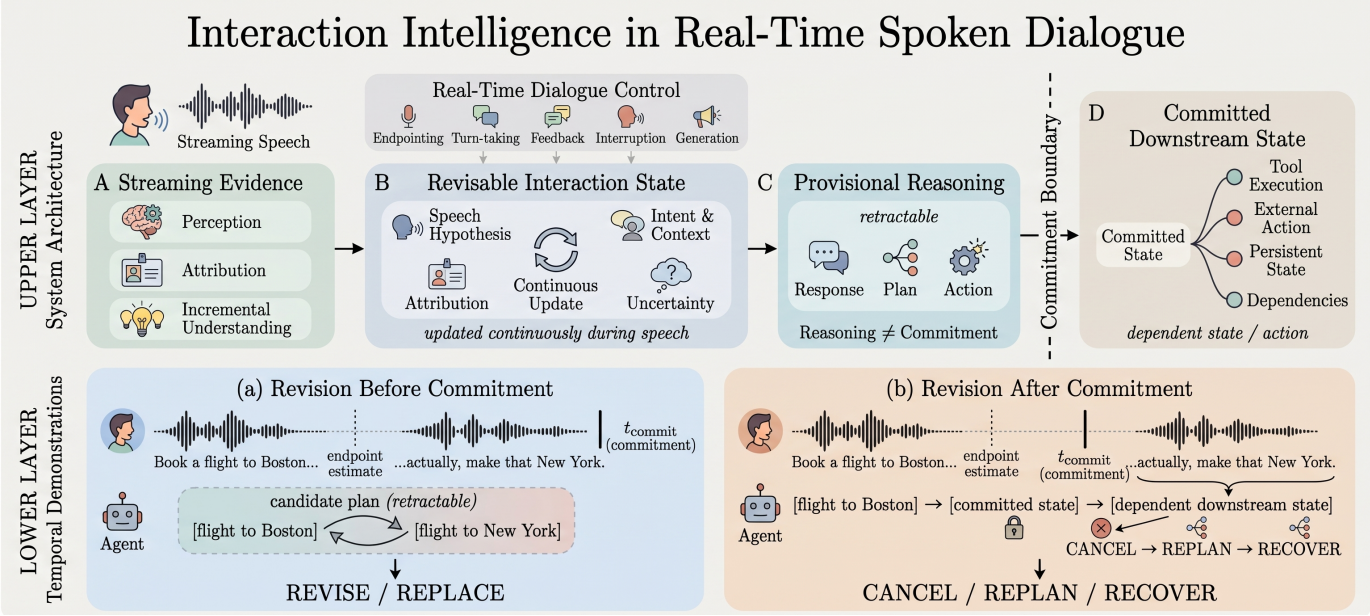

图 3. 实时口语对话中的交互智能概览。流式语音证据通过感知、归因和增量理解，持续更新可修订的交互状态。实时对话控制基于该状态运行，同时可在作出承诺之前暂时准备响应和计划。之后的口语纠正可能触发修复和重新规划，而越过承诺边界则会使下游状态产生实际影响。因此，停止语音输出与取消已准备的动作是不同的控制决策。

C. 理解与预测：端点与承诺

口语理解已不再局限于从单一最佳转录文本中提取意图和槽位。混淆网络保留了用于对话状态跟踪的备选识别假设 [24]；意图与槽位联合模型为流式输入分配标签 [13]；时间表达网络则表示信念状态在不同轮次间的变化 [25]。在有限领域中，意图准确率、槽位 F_1和联合目标准确率仍是有用的衡量指标，但它们并不能证明请求已得到妥善完成。即使正确理解了预期操作，该操作仍可能未经授权；而看似正确的槽位值，也可能被用于错误的动作。

另一个相关挑战是判断话语内容何时已足够完整，可以继续处理。基于停顿的端点检测必须在打断语速较慢的说话者与延迟响应之间作出权衡。轮次完成预测利用逐步获得的对话证据，预测说话者可能何时结束 [109]。语义端点检测将部分含义纳入这一决策 [14], ，而流式端点检测器可以通过延迟标签进行训练，以便在声学边界完全明确之前支持决策 [86]。更积极的预期会在轮次结束前预测其结束时间：据报道，在 Unmute 集成中，提前最多 2.56 秒作出的预测将平均响应延迟缩短了约 505 毫秒，同时产生了 28.4% 的推测计算开销 [85]。

尽管如此，端点估计不同于作出承诺的决定。我们用 t_{end} 表示估计的声学或语义端点，用 t_{commit} 表示承诺时间。前者关乎说话者是否已经说完；后者则确定系统何时允许当前解释影响持久状态或具有实际后果的下游行为。在 t_{end} 之前，系统可以进行推测性推理，前提是其中间状态仍可撤回。或者，系统也可以在端点之后继续等待，以解决说话者归因、歧义或授权问题。关键的设计选择在于，如何在早期解释与作出承诺之间处理不确定性。

每当说话者改变主意时，这一区别就至关重要。迟到的纠正可能会使待处理工具调用中的参数失效、改变计划，或要求重新考虑先前的状态更新。交互层必须能够将这一修订传递给下游，而不是把最新的转录文本视为孤立的替换内容。这里给出了一个具体的时序示例，说明图3中总结的区别。在 t_{commit} 之前，候选解释和计划仍可撤回，也可以直接修订；之后，其他组件可能已经依赖它们，因此需要取消、重新规划或恢复。

轮流发言长期以来一直是口语对话研究的核心，其控制策略也从启发式方法逐步发展为学习策略 [110]–[112]。人类对话轮次的时序为这些系统提供了重要约束 [113]。韵律和词汇线索有助于预测何时需要作出回应 [42] ，以及智能体何时可以开始语音表达 [43]。后来的模型结合了声学、词汇和上下文信息：一种多模态方法在真实的交互式语音应答语料库上报告了插话检测的 0.826 宏平均-F_1 [84], ，而一种四状态分类器则以约 95.75% 的平均准确率区分停顿、反馈语、打断和继续，其结果基于所报告的测试集 [46]。反馈语预测同样已从语音活动预测 [114]发展到声学与语言模型融合 [115]以及跨语言迁移研究 [116]。应答确认与打断之间的区别尤其重要：一个只要听到简短回应就停止语音表达的系统，可能在正确检测到语音的同时，却无法妥善管理对话

不佳。

插话使轮次协调重新回到归因问题上。当有人与智能体同时说话时，控制器必须判断这段语音是否是对系统说的、是否应停止其输出，以及新的话语是否改变了当前目标。设备指向性有助于回答第一个问题 [47]，而个性化活动检测和声学处理则可减少播放内容和干扰造成的误触发 [45], [94], [95]。停止音频输出只是可能的响应方式之一。如果被打断的话语与持续推理或已准备的操作相关联，控制器还必须确定这项工作能否继续、修改或取消。

插话是更广泛的重叠问题的一种形式。说话者可能同时发言，听者可能在不接过发言权的情况下表示认同，而纠正也可能在前一个话语结束之前就开始 [117], [118]。如今，帧同步模型、流式上下文和上下文感知基准将轮流发言视为一项声学与语言联合任务 [119]–[122]。其他研究考察打断的含义，而不仅仅是其时序 [123]–[125]；与此同时，抗干扰融合和用户角色控制则致力于提升系统在更复杂条件下的稳健性 [126], [127]。总的来说，这些研究结果表明，对话协调不能简化为检测麦克风通道是否处于活动状态。

D. 在交互循环中生成语音

语音生成为交互循环带来了另一组目标。识别通常通过转录错误来评估，而合成则必须兼顾可懂度、自然度、说话人身份和韵律保真度。Tacotron 首次实现了从文本到语音表征的端到端预测 [128]；Tacotron 2 将梅尔频谱图预测与 WaveNet 声码器相结合 [129]；HiFi-GAN 则提高了高保真波形生成的效率 [130]。随后，对话语音合成将对话历史纳入声学生成 [31]，而基于参考的韵律与节奏评估指标则提供了不再仅依赖平均意见分的替代方案 [131]。这些进展改善了用户听到的语音，但它们本身并不能决定系统应如何应对不断变化的对话。

声音还承载着情感和说话人身份。StyleTTS 2 将风格扩散与对抗训练及大型语音语言模型相结合，以提升自然度和说话人适配能力 [132]。研究人员已探索在合成和零样本语音转换中利用语音语言模型表征进行情感迁移 [133], [134]；也研究了具身场景中的长时程表现力控制 [135]；带有情感标注的对话语料则支持共情对话相关研究 [136]。其他方法将韵律与多智能体推理相结合，在响应生成前编码情感结构，或围绕冻结语言模型训练共情口语聊天机器人 [137]–[139]。这类研究让生成语音更具表现力，但合适的发声风格并不足以证明系统理解了情感，也不足以证明它能在高风险交互中安全行事。

验证也开始融入语音生成过程。CycleSpeech 通过结构化语音档案，将指令控制合成与副语言理解联系起来。在其前向循环中，系统会将从合成语音推断出的档案与预期的声音属性进行比较；在反向循环中，则利用从真实语音推断出的档案来重构原始说话风格。该方法使用 20,046 个双语示例进行训练，据报告，在中文和英文两种语言中的指令匹配准确率均高于其 Step-Audio-2-mini 基线 [140]。这里的验证针对的是语音信号及其预期属性，并不能证明某项外部操作已执行，也不能证明其结果正确。

语音翻译提供了一个相关视角，因为其评估一直围绕几个相互竞争的目标发展。级联系统保留了可逐阶段检查的中间表示，而直接方法则在没有显式转录文本的情况下生成目标语音 [32]–[36]。BLEU 和 ASR-BLEU 衡量内容保真度的某些方面；平均意见分和说话人相似度评估语音质量；实时因子或延迟则捕捉时序。同步翻译尤其清晰地展现了这种权衡：在源语音结束之前开始生成目标语音可以降低延迟，但可能会损害充分性。新的基准测试因此在传统内容指标之外，还考察表现力、长篇同步翻译以及多轮质量 [141]–[144]。多语种研究还进一步测试不同口音、低资源语言、发声模式和类型学条件下的性能 [101], [145], [146]。强劲的多语种总体表现可能掩盖语言之间的显著差异，而当智能体依赖识别或生成来完成任务时，这些差异会在实际操作中变得十分重要。

因此，与其从对话后果层面衡量，语音生成更适合在信号层面进行评估。自然度、保真度、翻译质量和延迟都已有成熟的评估程序。人们对生成语音如何影响协调与动作的了解则相对有限：例如，打断发生时，输出是否会适时停止；发声风格是否会改变用户对请求的解释；或者听起来充满自信的回答是否掩盖了一次未成功的操作。当听觉理解与语音表达同时发生时，这些问题变得更加迫切。

E. 全双工技术

全双工对话允许系统在自身语音播放期间接收用户音频，并决定继续、停止、让出对话主导权、发送反馈语或恢复发言。这涉及持续的交互控制，而不仅仅是同步翻译、较低的生成延迟，或单纯重现两路音频流。模型可能具备上述任何特性，却无法可靠地管理对话主导权。

对并发进行建模。双通道表述早于近期的语音基础模型。生成式口语对话建模表明，对话双方都可以表示为并行流，包括自发重叠 [39]。Moshi 通过并行的语义与声学编解码流以及内部文本

流进一步发展了这一方法，据报告，其理论延迟约为 160 毫秒，实际延迟为 200 毫秒 [1]。文本流被描述为内心独白，可作为中间表示，与连续声学生成并行。其他系统将决定倾听还是说话作为明确的学习目标：SALM-Duplex 学习何时倾听、何时说话 [6], , 而 Freeze-Omni 则围绕冻结语言模型构建双工语音接口[7]。GLM-4-Voice 和更长上下文的模型致力于端到端的对话语音 [48], [49]；SALMONN-omni 在没有编解码器瓶颈的情况下连接流式编码与合成 [11]；模块化替代方案则为原本传统的后端添加双工控制 [53]。在集成式设计中，继续还是让出发言权的选择本身成为生成的一部分。这有助于使时序决策与用于生成语音的表示相协调，但也可能使这些决策如何作出缺少更明确的证据。

基准测试及其带来的变化。评估也随这些交互日益复杂而不断发展。早期针对受控音频的轮流发言指标 [147] 已扩展至重叠处理 [54], 、自动化多轮评测 [55],, 以及与模拟工具 API 配对的含不流畅现象的人类录音[56]。互补性的基准测试在受控时序和噪声条件下生成高打断对话 [57],, 单独评估打断后恢复 [58], 、考察类人性 [148], [149], , 或评估流式语音到语音系统中的自然度 [150]。Talking Turns 为评估音频基础模型中的轮流发言行为提供了另一种工具 [151]。

这些基准测试越来越多地分别报告以下指标：接过话轮、打断、停止与恢复、恢复能力以及响应质量。Full-Duplex-Bench-v3 说明了进行这种区分的必要性。在来自 12 名说话者、涵盖四个模拟 API 领域的 100 段录音中，一个实时模型的工具选择 F_1 达到 0.876，pass@1 达到 0.600，接过话轮的比例为 96%，任务延迟为 6.89 秒。一个响应更快的竞争模型仅需 4.25 秒即可响应，但只有 78% 的情况下能够接过话轮；而级联系统在 10.12 秒时接过每一轮，且打断率更高 [56]。单一的延迟或成功指标会掩盖这些差异。

通过结构化动作实现双工控制。当结构化动作与语音共享同一时间线时，能动性便与之有了更直接的联系。DuplexSLA 在共同的 160 ms 时钟上同步用户音频、助手音频和动作输出。它会在语音持续进行的同时生成规划文本、JSON 调用以及中断或反馈语标签。在 900 个确定性工具案例中，其平均准确率为 85.56%，延迟为 0.648 秒；级联系统的平均准确率为 91.33%，延迟为 2.77 秒；不同状态下的轮次决策准确率则介于 93.33% 至 99.33% 之间 [20]。这表明，在基准测试的时序限制下，语音生成与结构化调用生成能够协同进行。其确定性的函数与参数匹配标准，并不能证明系统获得了授权、真正执行了服务，或验证了外部状态变更。

近期的一个具身系统让承诺的时序变得更加直观。MIRA 在 Astribot S1 机器人上结合了流式语音、轮次仲裁和伴随语音的动作。其 CORTEX 控制器将快速打断处理与审慎的轮次决策分开，而 ROSCO 使用滚动时域前缀承诺生成动作：它预测的范围比实际释放的更远，因此只有一小段可取消的动作前缀会进入物理执行 [152]。动作一经派发，就无法像内部文本计划那样自由修订。因此，限制已承诺的前缀，便提供了一种具体的可进行时间修订的控制方式。报告的实验涵盖了真实机器人流式交互和打断处理，但并未验证从授权到外部状态验证，再到补偿的完整链条。表 I 对比了有代表性的交互系统和基准测试；由于 MIRA 是在检索截止时间之后才出现的，因此在此单独讨论。

F. 修复：从停止到重新验证

系统可能及时检测到打断，却无法妥善处理其后果。在一项涵盖 45 段合成状态机对话、10 个企业领域、27 种模型配置和 6 种中断类型的评估中，任务完成率最高的系统与恢复质量通过率最高的系统并不相同。尽管自动评审之间的一致性很高，它们与人工评审的一致性仍然不高 [58]。其他研究发现，系统虽然经常能将重复请求识别为需要恢复的情况，却未必能始终如一地修复 [57],；或者模型会在用户发出自我纠正之前，就认定用户仍怀有过时意图 [56]。恢复对话、停止语音输出以及取消下游动作，是彼此不同的结果。

关于言语不流畅和语音修复的研究，为理解这类纠正提供了语言学基础 [153], [154]。然而，对于语音智能体而言，修复必须触及执行任务时所使用的状态。话语中途更正可能会改变目标、待处理调用的参数，或系统已经准备好的计划。对话策略中的澄清和智能体框架中的恢复处理了这一问题的部分方面 [155], [156],, 但所引用的评估分别考察了这些方面。因此，当前的中断基准测试更清楚地证明了系统是否停止或恢复，而非系统是否正确修订了预期操作。

G. 交互智能未能解决的问题

交互层的衡量方法相对成熟。识别、说话者归因、端点检测、状态跟踪和轮次控制，都可以通过专门的指标和已知失败案例进行评估。它们的输出也会影响交互层之外的决策。说话者归因在不同说话者拥有不同权限时至关重要；端点检测会影响暂定解释何时可用；而插话可能要求修订正在进行的计算或已准备的动作。这些上游决策的质量固然重要，但其本身并不能决定最终动作是否合法或正确。

全双工技术探讨的是，当人和系统都可以说话时，双方如何共享对话主导权。语音智能体还必须决定何时依赖某种解释、请求执行的操作是否经过授权，以及动作改变外部状态后该如何处理。这些职责需要明确的机制来管理状态、推理、记忆、执行、观测和恢复。下一节将探讨该智能体

的循环，以及它通过哪些接口接收来自语音交互的暂定证据。

IV. 智能体智能：从语音意图到具有实际影响的行动

当语音交互中的解释被保留下来、用于准备一项操作，或传递给能够改变持久状态或外部状态的系统时，它便会产生实际影响。推理和工具调用提供了重要能力，但单靠其中任何一项，都无法确保用户的目标已贯穿规划、执行、观测和恢复 [157], [158]。我们用智能体智能来描述语音智能体这一以目标为导向的维度。它无法与交互智能完全分离：计划可能取决于仍在修订中的语音，而采取行动的权限则取决于请求由谁提出以及提出时的情境。问题在于，当解释逐渐转化为决策，并最终产生可观测的后果时，如何保留这些依赖关系。表II总结了其中涉及的接口，以及由此产生的人类控制义务。

A. 状态与语音推理

在当前的语音智能体架构中，推理处于不同的位置。它可以表现为基于显式对话状态的策略、语音前端之后的语言模型，或直接处理声学表征的原生语音模型。其他设计则采用独立的规划器，或将请求委托给外部智能体。这些方案的差异不仅体现在灵活性和延迟上，也体现在决策出错时能够检查哪些内容。

经典对话系统使决策过程相对明确。Plato 将对话状态、策略和服务划分为可检查的组件 [23]；通过强化学习得到的策略基于跟踪的信念状态运行 [29], [30]；此外，交互质量估计也已被研究用作策略学习信号，而不只是事后评分[159]。采用显式状态表示后，可以区分错误的信念更新、糟糕的策略选择和未成功的服务操作。其权衡在于，固定表示可能难以适应被打断的请求、对先前约束的撤回，或依赖较长对话内容的指代。

将语音前端连接到语言模型推理器可提供更大的灵活性。音频模型可以由语言模型协调 [60], ，通过思维链训练可以鼓励中间推理 [15], ，而流式系统可以在话语尚未结束前，根据语义事件启动推理 [16]。然而，这种架构也会使不确定性不断传递。如果模型依赖转录文本，识别错误可能改变推断出的目标；如果模型生成了一个看似合理的计划，该计划仍需有证据证明其正确性。审慎推理也会增加延迟，因此，推理策略的选择会影响第 III 节讨论的交互预算[160]。

语音原生模型避免将转录文本设为感知与推理之间的必经接口 [11], [12]。一些双工架构还保留着内部文本流同时继续进行声学生成 [1]。这些设计能够利用韵律、背景声音和重叠语音，而这些信息可能无法保留在转录文本中。然而，报告中列出的评估关注的是不同问题。语音推理基准考察推理、口语和副语言理解 [161]，而面向工具的测试则包括监控和触发检测 [156]。单凭任何一类结果，都无法证明模型具备校准置信度，也无法证明其在声学证据不确定时能够恰当地拒答。

因此，架构选择决定了哪些决策证据得以保留。显式策略会暴露中间状态；集成模型可能使从感知到响应的路径更难检查；独立规划器则通过结构化输出恢复一定的可见性，但代价是增加额外的接口。无论采用哪种设计，重大决策都必须能够追溯到其所依据的信息。

B. 规划与重新规划

规划将经过理解的请求转化为行动方案。面向任务的框架会将目标分解为结构化操作，供执行器执行 [3], [8]。情境感知智能体还必须随着环境变化修订其行为。例如，DOROTHIE 将对话决策与导航结合起来，应用于驾驶模拟器中；在该模拟器中，智能体所说和所做的都必须响应同一不断变化的情境 [9]。

语音使计划格外容易受到修订影响。说话者可能会澄清指代、更改约束，或在某项操作准备过程中打断对话。对话策略研究探讨何时应寻求澄清 [155]，但只有当澄清结果传达到依赖计划时，澄清才有用。此处考察的研究在不同情境下探讨了澄清与恢复；它们对于从口头纠正到重新规划以及取消过时操作这一完整流程，提供的直接证据较为有限。

一旦提案进入执行阶段，这一流程就会变得更难管理。内部计划通常可以在不产生外部影响的情况下进行修改；而已经派发的操作则可能需要取消、采取补偿行动或进行人工干预。语音与动作同步的系统缩短了解释与准备好的调用之间的间隔 [20]。因此，区分候选操作与获准执行的操作尤为重要，并且必须使依赖已被替代的语音所作的决策失效。

C. 记忆与个性化

记忆使语音智能体能够利用当前话语之外的信息，尽管人们常用同一个名称来讨论几种不同的机制。交互上下文 保留理解当前轮次所需的近期交流内容。检索记忆从外部存储中提供信息，包括通过感知音频的检索或无转录检索 [70], [71]。持久自适应记忆 则将偏好、观测或事件片段跨会话保存，并将其与某个身份相关联 [2], [17]。最后一类记忆带来了特殊的责任：在一次对话中收集的信息，可能会在记录时的情境已经改变很久之后，仍然影响某个动作。

检索已有成熟的方法论基础。密集段落检索针对开放域问答学习段落检索 [162]，而 RETRO 则展示了由外部词元数据库支持的语言建模 [163]。面向语音的系统可以将知识保存在语音模型之外，以便独立更新 [72]，；也可以异步检索，而不阻塞正在进行的交流 [164], [165]，；还可以利用声学证据，而非单一的转录假设 [70]。回忆、回答准确性和忠实度描述了检索信息并据此回答问题的效果，但这些指标并不能证明检索到的偏好属于当前说话者、仍然有效，或可以正当地用于影响某个动作。

与其只考察检索，不如从条目的生命周期来审视持久记忆。一组有用的操作包括写入 → 检索 → 更新 → 整合→ 纠正/删除。写入决定哪些信息会被持久保存；检索将先前的信息引入决策；更新和整合则使新的观测与已有存储信息相协调。纠正和删除使用户能够修订或撤回信息，避免这些信息继续影响行为。在共享设备上，这些操作还需要明确的身份边界 [18]。注册、说话人验证和身份识别是不同的任务；识别出某个声音本身，并不意味着获得了访问相关记忆或据此采取行动的权限 [108]。

当记忆影响规划时，两种失效会产生尤为严重的后果。时间漂移 指随着音频上下文不断累积，推理、记忆或对话连贯性逐渐退化，从而限制了根据较短会话能够得出的结论 [73]。来源信息丢失 则发生在摘要生成或整合模糊了已存陈述的来源、时间或后续纠正时。任一情况都可能改变计划或工具参数，却没有留下清晰易懂的记录来说明变更原因。

因此，对于具有重要后果的用途，记忆条目应保留其身份范围、来源、时间背景和修订历史。设备端方法和隐私保护方法可以减少部分个人信息暴露 [77], [104]，而智能体框架则可以向其他组件公开生命周期事件[8]。难点在于将纠正传递到记忆存储之外，传达到任何依赖过时信息的待定决策中。

D. 检索与工具增强型语音智能体

工具为口头请求提供了通往外部服务的途径。选择函数并生成其参数是成熟的研究问题 [166], [167]。一个数据集包含三万个经过语音克隆且带有噪声的请求，涵盖智能汽车、智能家居和可穿戴设备领域，并包含单意图、多意图和大海捞针设置；声学扰动和组合性要求都会降低选择性能 [19]。实时评估纳入不流畅的人类语音，并报告工具选择 F_1 和参数准确率 [56]。其他框架则会检查生成的调用是否与其指定目标相符 [62], [168], [169]。

这些测试确立的是调用的形成，并不一定能证明授权、执行或成功。 当动作生成与语音表达共享同一时间线时，这一区分尤为重要。双工模型可以在用户仍在说话时生成规划文本、JSON 调用和轮次控制标签 [20]；纠正测试则考察说话者修改意图后，模型是否放弃了先前的意图 [56]。因此，在提出该调用的话语尚未完全明确之前，调用可能就已经准备就绪。最终回复正确，并不能证明已准备好的操作被安全地暂缓或修订。

基于真实环境的评估通过检查最终任务状态，提供了更有力的证据。一项涵盖 278 项零售、航空和电信任务的基准测试，结合了显式策略、可调用工具和目标数据库状态。对于原生音频服务提供商，该基准测试报告的 pass@1 在干净环境下为 31–51%，在真实环境下为 26–38%；其中相当一部分语音和噪声敏感型失败被归因于智能体行为，而非仅仅是声学前端 [63]。智能体框架还提供类型化执行器、安全组件和生命周期事件；据报告，一条规划器路径在 13 类请求中均达到其工作标准，而其他配置中的回退行为则较弱 [3], [8]。应用范围从农村表单填写和对话式搜索，到领域支持和多模态技能调用 [64]–[66]，[88]，包括临床预约安排和院前评估 [80], [170]。解读其报告的结果时，必须考虑具体的环境和任务定义。

商业案例描述了迈向实际运营应用的又一步。Yuma Voice AI 记录了这样的工作流：智能体识别来电者、检索订单信息、获取口头确认，并执行地址变更、取消、退款或订阅更新等后端操作。相关资料还介绍了转录文本和可配置的人工转接 [171], [172]。这表明供应商报告了一种将语音连接到具有副作用的服务的生产工作流，但并非对授权准确性、操作后验证、回滚或长期用户结果的独立评估。

即使有类型化执行器，可用操作范围仍会限制系统可以尝试的事项，而格式正确的调用也可能未经授权或遭到拒绝 [8]。真实服务还会出现超时、格式错误的响应、并发变更和部分失败，而确定性模拟工具无法完全捕捉这些情况 [56]。因此，评估工具使用时，必须区分发出调用、调度调用，以及确认调用之后发生了什么。

E. 数字与物理行动

物理动作让口语解读所带来的后果更难被忽视。PaLM-E 将具身观测与语言建模相结合 [173]；RT-1 基于真实世界数据学习机器人控制 [174]；RT-2 则将视觉语言表征与机器人动作联系起来 [175]。面向语音的系统还要应对对话中的时序与歧义。DOROTHIE 将对话与仿真中的导航联系起来 [9]；其他系统则利用多模态上下文进行主动操作，或实现低延迟服务机器人交互 [21], [89]。自我中心视角助手必须根据佩戴者视角解析指代 [22]，，而口语人机协作则依赖于建立共享上下文 [176], [177]。

同时进行语音表达和行动可能导致语音与动作处理之间的竞争；有报告指出，一种

系统在两者同时运行时，长时域任务性能会降低 [4]。MIRA 通过轮次仲裁、插话处理和流式运动生成来解决相关的时序问题。其控制器只释放一小段已确认的运动前缀，并在机器人端进行关节和碰撞检查，同时允许后续预测的运动进行修订[152]。这限制了在口语证据仍属暂定时已提交的物理行为范围，但并未确立独立的结果验证机制，也未提供失败后的补偿。

在具身场景中，语音交互的若干常见特性会转化为控制要求。

指称落地。例如“左边那个”这样的指令，指向的场景可能会在机器人采取动作之前发生变化。因此，情境系统必须持续维护指称对象，并在必要时对其进行修订 [9], [22]。

时序。语音会在控制进行过程中到达。停止命令可能需要在当前轨迹完成之前中断运动，因此，端点检测和双工控制不仅关系到响应能力，也关系到安全性 [89]。

说话人身份。在共享空间中，识别出是谁说了话，并不足以确定此人是否有权发出特定命令。第III节讨论的归因机制必须与授权策略相衔接。

纠正与确认。修订可能会在运动开始后才到达。与候选计划不同，物理变化未必能通过编辑记录来撤销；停止、观察结果，并在可能的情况下进行补偿，都会成为控制问题的一部分 [4]。

人工接管。 在人附近运行的机器人需要有明确的方式交还控制权。辅助场景中的注意力与听觉理解研究会考察接管提示和控制权交还行为 [178]。

行动中的恢复。抓取失败、遇到意外障碍物或指令被打断时，可能需要在下一轮对话开始前作出响应。因此，必须在控制仍处于激活状态时评估恢复能力。

这些要求的证据来自不同的实验场景。Habitat 为具身仿真提供了平台 [179], ，而 ALFRED 则评估语言引导的家庭任务 [180]。驾驶模拟器报告受控交互条件下的导航结果 [9]；实验室研究考察较短会话中的自然度、适切性和接管情况 [178]；监督式工作流程则提供范围更窄的部署证据 [80]。仿真可以模拟特定干扰因素，但无法重现部署中遇到的各种声学、感知和运行条件。短期实验室研究也无法充分说明持久状态随时间的变化情况 [22], [181]。

其他研究将语音与视觉上下文相结合 [182]–[184],，将推理和工具引入口语对话 [185], [186],，并评估边界意识或多步骤辅助，而非仅评估响应质量 [187]–[189]。随着这些进展不断推进，观察动作所产生的影响，而不仅仅是智能体提出动作的能力，变得愈发重要。

F. 行动闭环：从调用到验证后果

有效的工具调用并不能证明预期的改变已经发生。我们将行动闭环 用于描述后果侧的循环执行 → 观察 → 验证 → 恢复或补偿。这是第一节 B 小节中引入的更广泛的能动性闭环的一部分，其中还包括语音交互、授权承诺，以及向用户传达经验证的结果。

这些阶段分别回答不同的问题。执行关注操作是否已被派发，以及服务或控制器如何处理该操作。观测记录尝试后报告的状态。验证将该状态与预期结果进行比较。如果两者不同，恢复可能需要重试、重新规划、补偿或升级处置。选定的函数可能从未被调用；已接受的请求可能只产生部分变更；而报告的状态可能无法满足用户的目标。单一的成功标签会掩盖这些差异，也无法明确指出需要在哪里进行干预。

图4将这一循环呈现为责任与证据的映射。语音输入、记忆和环境上下文会为决策提供信息，但提出、授权和执行一项操作在逻辑上仍彼此 distinct，即使它们共享同一实现。三个时间点对于遏制错误尤其有用：派发前、观测结果后，以及传达成功之前。

现有评估揭示了这一循环中的不同部分。语音到函数和Duplex基准测试会在噪声、言语不流畅或打断条件下测试函数选择和参数生成 [19], [56]；智能体框架会体现类型化执行和结构化故障处理 [3], [8]；基于落地任务的基准测试会将最终数据库状态与指定目标进行比较 [63]；而中断研究则将恢复质量与任务完成区分开来 [58]。每项评估都在特定环节提供证据，但仅凭其中任何一项调用级指标，都无法确立经过验证的后果。

这些是逻辑职责，不一定对应彼此独立的软件模块。规划器提出操作，授权机制决定是否可以继续，执行器执行该操作，而验证则评估观测到的结果。只要相关转换和证据仍然可访问，同一个模型或运行时可以实现多个角色。否则，流畅的语音确认可能表达的只是系统预期会发生什么，而不是它已确认实际发生了什么。

同样的边界也决定了何时必须进行人工控制。能够改变持久状态或外部状态的智能体必须明确谁可以授权该变更、它可以使用哪些信息，以及当结果错误或不确定时用户如何进行干预。这些问题促成了下文将讨论的治理和人工控制机制。

V. 以人为本且值得信赖的语音智能体

可信赖性是使语音智能体产生实际影响的同一组状态转换的一部分。身份决定哪些证据会进入循环；授权限制哪些计划可以执行；隐私和同意约束持久状态；可及性决定谁能可靠地使用交互；验证和恢复则限制一次

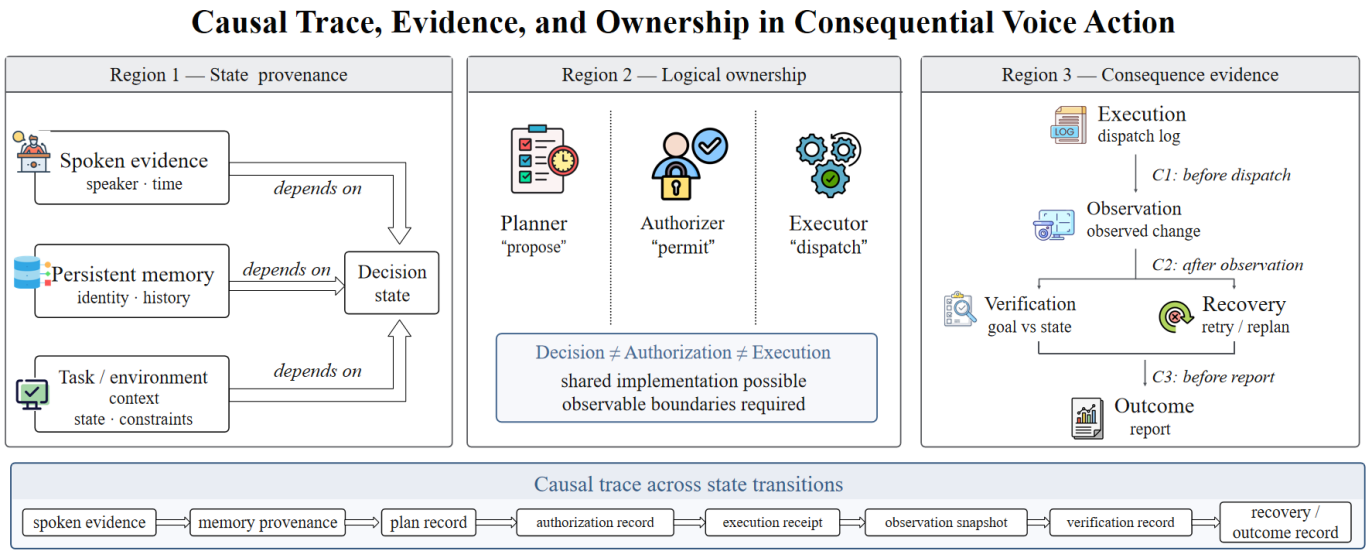

图4. 重大语音操作中的因果轨迹、证据与所有权。口语证据、持久记忆和任务上下文为决策状态提供信息。规划、授权和执行具有不同的逻辑作用，而观测、验证和恢复则提供结果证据。底部色带追踪这些依赖关系；C1–C3 标示了在派发前、观测后以及报告成功前遏制错误的机会。

错误动作的影响。这些既是架构约束，也是部署策略。图 4 总结的所有权和证据边界，为探讨身份、权限、隐私、可及性和监督如何约束会产生实际影响的状态转换，提供了一个共同框架。

A. 个性化、身份与授权

个性化将持久状态引入这一循环，而持久状态也伴随着相应的义务。写入偏好、稍后检索该偏好，以及据此采取行动，是三个不同的决策。类别边界记忆通过对关于用户的长期信息进行组织，来控制无界增长 [17]，；流式双存储记忆则让检索能够在交流过程中进行，而非只能在轮次之间进行 [2]。这两种设计都没有回答用户有权提出的问题：关于我的信息存储了什么、为什么存储，以及我该如何将其删除。偏好命中率指标衡量的是实用性，而非控制权。

身份是一项独立的技术。相关研究将注册、验证和身份识别区分为不同任务 [108]。正是这种区分使授权得以明确表达。系统可以识别出某个声音，但仍须确认此人是否有权发出特定指令；身份感知记忆也表明，这种失效模式确实存在：在共享设备上，除非预先设计隔离机制，否则不同说话者之间可能发生上下文泄漏 [18]。识别只能作为说话者身份的证据；它并不意味着获准执行动作。欺骗攻击对策也与身份凭证的可靠性相关：ASVspoof 2019 评估了对欺骗语音和合成音频的检测能力 [190]。

授权是一种属性，它将技术上有效的动作转化为获准执行的动作，并以一种特定方式约束系统设计：负责决策的组件与负责执行的组件必须由一道承载身份和范围的边界分隔开来。面向使用工具的语音助手的治理框架已经开始规定完整性、访问控制和审计要求，并且表明这些限制会改变所测得的泄漏情况 [78]。现有评估尚未能在语音交互中同时纳入被拒绝、过期、含糊不清和已撤销的权限，并同时衡量不安全操作与正确的拒答。在缺乏这类评估的情况下，工具选择准确率和通过率不能替代授权测量。

B. 同意、隐私与安全

语音智能体中的同意并非注册时的一次性事件，而是贯穿采集、保留、再利用和删除的生命周期。保护隐私的端侧助手表明，将音频保留在本地可以降低部分相关风险 [104],；交互隐私基准则衡量语音模型会透露对话中的哪些内容，而不只是转录了什么 [77]。两者都是对接口的测量：前者关注计算发生的位置，后者关注模型会透露什么。模型记忆还会带来一种独特的隐私风险：训练数据提取研究表明，语言模型可能泄露部分训练数据 [191]。

安全性研究结果与架构尤其相关，因为持久性和行动扩大了攻击面。早期针对语音转文字系统的攻击展示了定向对抗性转录 [192] ，以及如何利用心理声学约束隐藏扰动 [193]。语音特定对抗性输入可以通过不可感知的扰动劫持音频语言模型 [75], ，并且已有防御措施被提出，用于保护语音语言模型免受越狱尝试 [76]。更广泛的音频安全工具涵盖多种威胁模型[194]。持久且具备行动能力的智能体还会暴露出另外两种故障模式。说话人反演表明，即使转录文本看似无害，语音令牌仍可能泄露身份相关信息 [195], ；拒绝服务攻击则表明，对抗性输入可以让智能体一直说话，从而无法采取行动 [196]。能够记忆并调用工具的系统可能通过其记忆和工具遭受攻击，而不仅仅是通过其识别器 [197], ，相应的

设计后果是，授权、内存写入和工具调用都各自需要独立的策略和审计记录。

测量基础设施也在同步发展。对多说话人语音中的操纵行为进行检测、通过反事实审计检验副语言证据是否确实被使用，以及在人工监督下研究评审者可靠性，这些工作都在探讨一个问题：所声称的能力是否就是实际测量的能力 [198]–[200]。

C. 信任、共情与互动质量

面向人的证据主要集中于专注倾听和具身对话。针对老年人的研究报告了回答被评判为适切的频率，同时也显示出相当一部分回答被评判为不适切；这使行为的局限性得以显现，而不是被汇总结果掩盖 [201]。表达性虚拟辅导系统分阶段记录其架构和评估过程 [202], [203],情感标注语料库为投诉处理和听觉理解提供对话数据[136], 而近期系统则将共情进一步融入对话模型本身，包括全双工共情模型，以及使用冻结语言模型进行的共情训练 [139], [204]。

时序、反馈语、韵律和具身性会改变用户对智能体的解读，也可能影响用户愿意说出或授权的内容。证据不应被延伸到这些终点之外。感知共情并不能证明情感理解，表达性输出也不能证明具备安全处理能力，而短时会话中的适切性评分无法预测长期的关系结果。因此，信任需要单独进行测量，包括开展随时间变化的测量 [205], [206]。

可用性证据也呈现出同样的模式。对现有工作流中语音智能体的研究报告了可用性评分，却未接入真实的记录系统，也未确立结果 [81]；对话式语音日记的实地评估衡量了依从性和自我报告的丰富程度 [207]；关于老年人对助手解释的感知的研究表明，个性和解释风格会影响接受度 [208]；课堂研究则表明，感知到的口音会影响用户与语音系统协作的方式 [209]。当系统是一种工具时，可用性是合理的评价指标；但若将其用来论证安全性，就会造成代理指标错误：高可用性评分只能证明人们能够使用该系统，而不能证明该系统做出了良好的决策。

D. 可及性、公平性与高风险应用

公平性证据体现为对语言、口音、年龄和残障覆盖情况的考察；从技术角度看，这一点至关重要，因为识别和合成质量决定了一个人能否使用智能体。自动语音识别中有记录的种族差异，具体说明了不同用户群体的识别性能可能存在系统性差异 [210]。多口音基准测试集、语码转换评估、抗口音识别，以及面向低资源变体的对话系统，揭示了哪些人群面临被排斥的风险 [98], [100]–[102]。特定领域识别失败集中在错误会造成严重后果的场景中：语码转换临床语音即使整体识别看似可以接受，语码转换语音仍会产生大量任务关键字词错误 [99]。不同人群的接受度也会有所差异，且这种差异独立于其底层识别性能 [208], [209]。本文考察的研究中，关于交叉性和纵向公平性的证据有限。上述证据按语言、口音、年龄和使用场景进行了细分，但几乎不足以支持关于这些因素交叉影响或长期结果的结论。

高风险应用使协助与权限之间的区别变得具体。结构化院前评估在临床工作流中使用语音智能体，并由临床医生确认 [80]；处方和会诊系统会调整语音模型，使其适应医学语言 [211]；相关综述梳理了语音和语言干预在心理健康护理中的应用场景 [212]。对于与认知障碍者收集多模态对话所涉及的同意与伦理问题，研究者已明确将其作为研究设计问题加以处理 [213]。在这些情境中，相关评价指标包括临床确认、升级处置、假阴性率和专业监督；对话质量并不能说明其中任何一项。本文讨论的健康相关结果，最好被理解为有关工作流适配性的证据，而非医疗疗效的证据。

E. 治理作为闭环属性

持久化且具备行动能力的智能体会以机械方式将这些治理问题耦合起来。从生命周期视角审视机器学习带来的危害，有助于区分源自数据、模型开发、评估和部署的风险 [214]。当系统向错误的人披露信息时，错误的说话者归因就会演变为隐私泄露。当呼叫已经准备就绪时，被误解的纠正就会变成未经授权的行动。当系统请求权限时，具有说服力的声音会放大同意问题。无法删除的记忆会削弱注册时作出的保证。不可逆的物理动作则会移除确认本应提供的恢复路径。

因此，在图 4 所概括的所有权和证据边界上，系统必须回答一些具体问题：是谁在说话；这段话语是否是对智能体说的；说话者是否获得授权；用户是否表示同意；实际执行了什么动作；有哪些证据能够确立结果状态；该动作能否撤销；以及智能体事后会如何告知用户。这些条件决定了能动性闭环中的承诺和后果阶段是否可审计：无法归因于某位说话者、无法依据某一身份获得授权，或无法在事后作出解释的行动，只是在软件意义上完成了闭环，而没有在人类意义上完成闭环。表 II 将这些控制措施与其所约束的能力接口并列呈现。这种并置是有意为之的：能够证明系统可以记忆、规划或执行的证据，并不能证明身份、同意、授权或升级处置得到了正确处理。因此，评估部分将这些控制措施视为后果性系统必须满足的可观测要求。

VI. 评估：让证据与能动性主张相匹配

评估让听觉智能与听觉能动性之间的区别具体化。识别、合成、

表 II 按所暴露的接口整理的智能体能力与治理约束。将能力证据与以人为中心的约束一并展示，是因为后果性行为既取决于系统能做什么，也取决于允许并可观察到这种状态转变的条件。

接口	机制或系统问题	有用的端点／时间范围	边界与代表性证据
A. 状态、推理与记忆			
显式状态 / 策略	符号式或学习得到的对话状态，以及一个针对已声明变量的策略	状态准确率；策略回报；任务成功 / 轮次—会话	使决策状态可供检查 [23], [24], [29]
流式推理	部分转录文本、语义事件或潜在语音状态触发审慎推理	触发质量；推理质量；延迟 / 轮次—会话	推理可以在一个轮次结束之前开始，但早期假设仍是暂定的 [15], [16], [160]
澄清 / 重新规划	不确定性感知问题、修订或放弃暂定计划	澄清准确率；修订后的任务结果 / 任务	将修订计划与仅仅生成另一个响应 [155], [156]
检索记忆	音频或基于文本的检索融入当前交互	回忆；回答准确性；忠实度 / 轮次—会话	检索质量并不能证明持久用户状态 [70]–[72]
持久记忆	偏好记忆、情景记忆或双存储记忆跨会话	偏好使用；身份隔离；长期时间范围质量／跨会话	持久性带来纠正、删除、来源方面的义务，以及漂移义务 [2], [17], [18]
B. 行动与后果			
语音到工具映射	函数选择与结构化参数根据语音生成	选择与参数准确率 / 轮	有效调用生成并不等于执行 [19]
同步工具使用	生成的动作令牌或结构化调用在双工语音时钟上	工具准确率；pass@1；延迟／轮次–task	时序证据仍有别于授权和副作用 [20], [56]
类型化执行 / 恢复	规划器派发受约束操作并重试或回退	完成；生命周期事件；回退行为 / 任务	执行过程可观察，但要确认成功仍需要结果验证 [3], [8]
基于依据的状态变更	针对目标评估的工作流环境或数据库状态	已达成状态与目标状态／任务	在所述环境中确立后果 [63]
具身动作	与导航耦合的对话状态，操作或情境控制	任务结果；控制成功；接管 / 回合	物理后果提高了可逆性和手-动恢复要求 [4], [9], [21]
C. 人工控制、安全与部署			
个性化 / 记忆控制	哪些用户上下文可以持续保留，以及如何可以纠正或删除吗?	偏好使用；纠正；删除；漂移	有用记忆不等同于可控记忆 [2], [17], [18]
身份 / 授权	谁在语音表达、服务对象是谁，以及这位说话者是否可以请求执行此动作?	验证；不安全操作率；正确的拒答	识别是身份凭证，而非行动许可执行 [8], [78], [108]
同意 / 隐私	哪些内容可以被采集、保留、再利用或披露?	同意生命周期；披露；重构风险	隐私约束随状态转移，而不仅仅随麦克风 [77], [104], [195]
安全 / 可用性	对抗性音频、记忆或工具输入会改变或阻断该循环吗?	攻击成功率；防御率；泄漏；可用性	持久性和动作扩大了攻击面 [75], [76], [196]
信任 / 交互质量	时序、韵律、解释和恢复能否支持校准依赖?	适切性；重复交互依赖	感知质量不能替代安全或正确的行动 [139], [201], [202]
可及性 / 公平性	谁会经历识别质量下降、合成或互动质量问题?	不同口音群体之间的差异，语言、年龄和残障	是否能够进入闭环，本身就是一种系统约束 [101], [208], [209]
高风险监督	哪些决策需要确认、升级处置、接管或专业判断?	升级处置；接管成功；已确认结果	工作流适配性并不能证明领域效能 [80], [81], [178]

延迟、轮流发言、检索和任务完成都有成熟的衡量指标，但在这些指标上取得成功，并不一定能证明语音智能体产生了预期后果。系统可能正确识别请求、推断预期任务并生成有效的工具调用，却代表错误的说话者采取行动、在收到纠正前就继续执行、修改错误的状态，或未检查结果便报告成功。因此，评估听觉能动性需要提供证据，将语音输入及其解释与承诺、经授权执行以及随后观察到的后果联系起来。

A. 从组件性能到后果

语音系统传统上通过针对特定能力设计的工具进行评估。图5对比了这种以组件为中心的做法与追踪交互及其后果的评估。为阐明不同实验能够证明什么，我们区分了六种常见的证据终点。

在*信号* 层面，识别、说话人分离、感知质量和延迟刻画了声学通道。*语义* 层面关注系统是否能还原预期含义，而*交互* 层面则考察时序、重叠、打断、归因和修复。这些测量共同描述了系统参与口语交流的可靠程度。但它们本身并不能说明，当交流促成某个动作时会发生什么。

其余端点关注的是进一步的步骤。任务评估会考察系统是否在既定工作流或测试环境中达成指定目标；如果基准测试只检查拟议的工具调用，其证据范围相应地仅限于调用的构建。后果评估会考察由此产生的外部状态，以及预期变化是否能够被观察和验证。人类结果评估则会考察这种行为如何影响相关人员，包括系统是否能在面对纠正、失败或重新获得控制权的请求时作出恰当回应。这些端点并非对系统进行普遍排名，而是描述支持某项具体主张的证据所能达到的范围。

这些区别在边界处最为重要。格式正确的工具调用可能从未执行。数据库基准测试或许能在模拟器中验证最终状态，却无法证明在线服务的可靠性。即使状态变化已经得到验证，也可能

如果请求者身份不当，或系统缺乏授权，这种变化仍可能不合适。反过来，端点检测研究并不一定要衡量人类结果，才能作出有效贡献。关键是让结论不超出实际观察到的范围。

B. 现有基准测试实际揭示了什么

表 III 汇总了这些端点上的代表性测量结果。这些结果保留了各自原有的实验场景，不应被解读为统一的性能排名。

各基准测试对成功的定义各不相同。流式理解评估的是话语结束前的意图和槽位预测 [13]。轮流发言和插话基准测试考察对话时序下的协调能力 [46], [84],，而双工评估则引入重叠、言语不流畅、打断和自我纠正[56], [57]。语音到工具基准测试通过测量结构化调用的生成，更进一步地考察动作 [19], [20]。基于环境的任务可以将由此产生的数据库状态与指定目标进行比较 [63], ；恢复基准测试则考察打断或失败后的行为，而非将常规任务完成视为充分条件 [58]。每项基准测试都扩展了流程中可观测的部分，但没有一项能让其他端点变得可互换。

近期系统进一步说明，“验证”一词需要有明确的验证对象。CycleSpeech 会检查合成语音是否体现了预期的声音属性。MIRA 报告了真实机器人中断处理的情况，并限制了计划动作中提交执行的部分。Yuma 的产品文档描述了能够触发后端状态变更的实时客服通话 [140], [152], [171]。这些分别是关于语音生成、具身控制和所报告生产工作流的证据。不应将它们视为对能动性闭环进行独立验证的等价演示；尤其是最后一个例子，其依据是供应商文档，而非独立评估。

受控基准测试同样会忽略系统开始执行动作后才会显现的重要条件。提供转录文本可以绕过识别不确定性。确定性模拟工具可能不会涵盖被拒绝的请求、超时、部分写入和并发状态变更。合成用户可能无法复现人们修改指令、撤回同意或纠正已经在准备中的动作的方式。这类简化有助于隔离机制。只有当在这些条件下获得的结果被用于支持更广泛的部署声明时，它们才会成为一种局限。

时间跨度形成了另一种边界。端点检测和工具调用测试往往在几秒或几分钟内完成，而持久记忆、身份隔离和恰当依赖则会在反复交互中显现。关于长时对话音频的研究已报告，随着上下文不断累积，推理、记忆和自我连贯性会出现退化[73]。因此，对持久自主性的评估必须持续足够长的时间，才能观察其声称要衡量的状态和行为。

C. 指标由其边界定义

即使是熟悉的指标，当其事件边界不同时，也可能难以解读。“延迟”可能指模型分块的延迟、首段音频时间、从估计端点到首次响应的间隔，或从麦克风输入到工具操作的时间。即使完成一项交互式任务需要几秒钟，同步动作生成也可能报告亚秒级延迟 [20], [56]。这些测量并不矛盾，而是描述了不同的时间区间。只有明确指定起始事件和观测终点，比较才有意义。

自动评审和汇总成功率同样需要谨慎对待。语言模型评审使大规模评估恢复能力、自然度和任务质量成为可能，但它们之间的一致性本身也是测量的一部分。在所报告的恢复实验中，自动评审之间的一致性超过了评审与人工评审员之间的一致性[58]。因此，当结论取决于这些分数时，校准和有针对性的人工评估十分重要。同样，单一的任务通过率可能掩盖失败究竟源于识别、归因、承诺时机、执行还是恢复。将失败追溯至特定接口的研究 [57],[63]能够提供仅凭汇总分数无法获得的信息。

D. 作为架构原则的闭环评估

在当前许多系统中，架构集成所涵盖的范围超出了评估记录的内容。语音智能体可能结合流式语音、推理、记忆、工具和恢复机制，但其各个组件却通过彼此独立的测试进行评估。双工系统可以在语音过程中生成结构化动作 [20] ； 智能体框架提供类型化执行和生命周期事件 [3], [8] ；基于真实环境的系统可以检查是否达到了目标数据库状态 [63]。更难从这些结果中重构的是连接这些能力的决策与观测序列。

闭环评估并不意味着要构建一个涵盖所有能力的单一基准测试。所需的追踪链取决于具体主张。对于高影响语音任务，它可能包括口头请求、说话者归因、暂定解释及其修订、承诺事件、授权、拟议操作与已下发操作、观测到的操作后状态，以及用于验证结果的证据。如果验证失败，追踪链还应显示系统是重试、采取补偿措施，还是将控制权移交回去。依赖记忆的任务还需要记录决策所使用的存储信息的来源；具身任务则需要记录对物理环境的观测以及控制权移交。不同的基准测试可以检查这一过程的不同部分，只要与其主张相关的转换过程保持可检查即可。例如，AgentDojo 提供了一个以安全为重点的评估环境，用于评估工具使用型智能体中的提示注入攻击及其防御 [215]。

这一要求对架构和测量都有影响。审计依赖于对受审查的决策和状态转换的记录 [216]。一个能够自主决策、执行动作并宣告成功的不透明组件可以

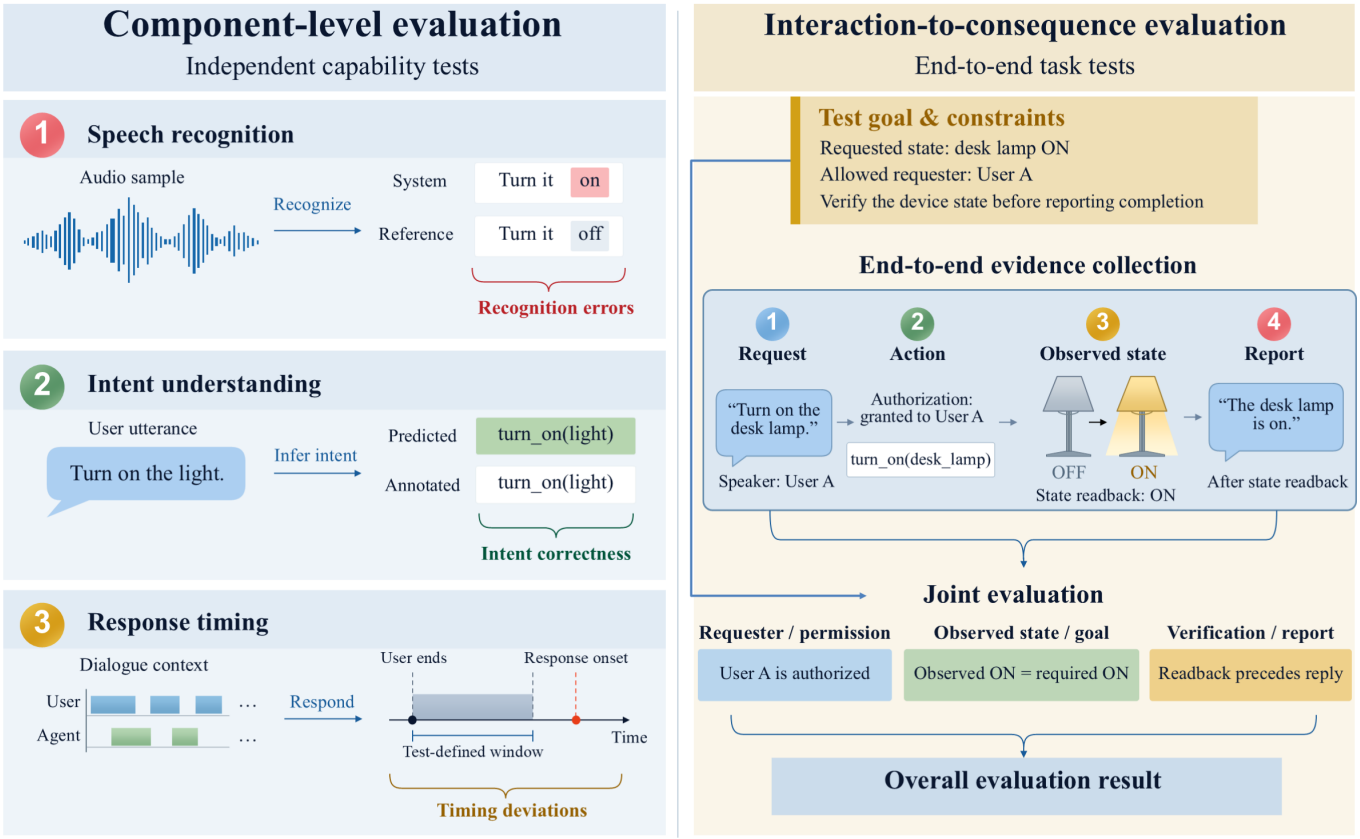

图 5. 语音智能体的评估范式 智能体。该图对比了当前的组件级范式与端到端的交互到后果
听觉能动性所需的评估， 并总结使此类主张可进行检验的评估设置和指标。

表 III 语音智能体的代表性证据端点。端点决定了所支持主张的范围。报告值保留其原始实验含义，且在各行之间不可直接比较。

端点	代表性证据	报告的结果	证据所证明的内容
信号	在所述测试条件下的四状态轮次分类条件 [46]	约 95.75% 的平均准确率	对停顿、反馈语的区分，打断和继续
语义	闭环中的流式意图和槽位定位设置 [13]	约 98.97% 的联合准确率	可在一个轮次结束前定位含义
交互	真实交互式语音应答音频上的插话预测 [84]	macro-F_1 0.826	系统应让出控制权时；下游状态未经测试
Task	900 个案例中的同步结构化操作 [20]	工具准确率为 85.56% 耗时 0.648 秒	低延迟调用生成；执行与授权仍然彼此独立
后果	278 项具有黄金数据库 状态的基于事实的服务任务 [63]	31–51% 清晰；26–38% 的真实感	将达到的模拟状态与目标进行核对
人类结果	与老年人的专注倾听互动 [201]	约 60% 的人认为合适	短时交互中的面向人类的适切性交互

仍应根据独立观测到的结果进行评估，但其内部故障可能难以定位。将决策责任与验证责任在逻辑上区分开来会有所帮助，即使它们由同一个模型实现也不例外，前提是支持各自的证据仍然可获取且可独立核查。模型自行声称某个动作已成功，不能替代对结果状态的观测。

组件级评估仍不可或缺。语音质量很重要，因为输入中的错误可能会传递到计划中；互动质量也很重要，因为不合时宜的

承诺可能会固守用户已经修正过的解释；任务能力也很重要，因为即使行动获得了适当授权，也可能失败。当一项主张跨越这些边界时，闭环评估会将这些测量联系起来。一旦语音智能体改变了外部状态，仅有决策证据便不再足够：评估还必须考虑之后发生了什么。

七、迈向通用听觉智能体

听觉能动性领域仍面临的挑战，出现在各项独立开发的能力相互衔接之处。流式模型

依托可修订的解释运作，记忆将先前的交互带入后续决策，而规划器和工具则将决策转化为持久影响或外部影响。纠正可能在任何时刻到来。系统必须在这些转换过程中保留用户的意图、动作背后的权限，以及其结果的证据。

四个接口问题构成了议程。表 IV 将它们作为实验来呈现，而非成熟度等级。

A. 可修正口语证据下的推理

响应迅速的语音智能体可以在说话者尚未说完时便开始处理。语义端点检测和预期轮次预测表明，部分语音足以支持时序决策[14], [85], ，而增量式系统无需等待完整话语[16]，即可根据语义事件启动推理。用户研究则从另一个角度提供了有关增量处理对交互影响的证据 [217]。更棘手的情况是，推理已经开始后才出现纠正。

暂定计算需要与已提交状态保持区分。早期假设可以用于指导检索、工具预取或候选计划，但不能据此授权不可逆操作。影响持久状态或外部服务的决策，需要有证据记录，并且能够修订依赖这些决策的其他决策。延迟训练端点检测器明确了早期决策的时序 [86], ；分阶段推理可以将部分计算从响应关键路径上移开 [160],；自我纠正场景则检验系统能否在用户更改请求时放弃过时意图 [56]。这些研究未必会追踪修订对外部状态的影响。

推测的边界可以直接检验：用户开始提出请求，智能体拟定计划，随后一句话使某项约束失效。评估将追踪哪些内容被修改或取消，以及在纠正生效之前是否已有外部操作执行。这将早期推理的延迟优势与正确性联系起来。

B. 身份、授权与持久状态

识别说话人只是决定是否根据口头请求采取行动的一部分。说话人验证、说话人分离、设备指向性和受话人检测，分别提供有关身份和对话意图的不同证据 [47], [105], [108]。授权还需要作出另一项判断：在当前条件下，此人是否有权请求执行特定操作。在共享设备上，同一请求可能因说话人不同而产生不同后果。

记忆会使这一问题延伸至不同时间点。跨会话和双存储方法会保留偏好与交互历史 [2], [17], ，而身份感知设计则力求确保这些记录与相应用户关联 [18]。对于执行型智能体，记录还需要包含身份、时间、来源，以及纠正或删除途径。

将记忆视为带有来源的状态，可以让规划器评估某项偏好对于拟议的动作是否仍然有效且适用。如果用户说“我不再想要那个了”，这项纠正可能需要使派生偏好、待执行计划或对原始记录的未来使用失效。关于记忆退化以及自我连贯性如何随交互累积而变化的证据 [73]，进一步说明有必要追踪存储信息如何变化，而不只是确认能否检索到这些信息。

过时的偏好、摘要漂移、用户间污染和不完全删除，可能只有在反复交互后才会显现。因此，纵向实验应在多个会话中跟踪用户及单条记忆记录，考察修订如何影响后续决策，以及过时信息是否仍会影响动作。

C. 闭合行动闭环

面向工具的语音系统如今开始展现的不只是最终的语音回答。语音到函数数据集会在噪声和组合性要求下评估调用选择 [19]；双工模型可以在话语尚未结束时准备结构化调用 [20]；基于实际状态的基准测试则会将最终数据库状态与指定目标进行比较 [63]。这些是行动路径上的不同环节。要将它们连接起来，就需要一个能明确区分授权、执行、观测、验证和恢复的控制流程。

这并不要求采用完全模块化的架构。规划器可以提出一项操作，但是否授权执行则是另一项决策。执行器可以报告服务的响应，但无法据此判断用户的目标是否已实现；这一判断取决于对观测状态与预期状态的比较。而恢复则取决于比较结果揭示了什么。即使一个模型承担多种职责，这些环节之间的转换也必须保持可观测。

在线服务会引入确定性模拟接口通常排除在外的故障：权限变更、过期记录、超时、部分写入、并发更新，以及无法撤销的操作。因此，评估除了奖励成功执行，也应奖励正确地不执行操作 。拒绝未经授权的请求、在收到纠正后取消已准备的操作，或在无法验证结果时升级处理，都可能是恰当的结果。在部分写入后进行补偿，则检验了更强的能力。现有恢复基准测试将修复质量与常规任务成功区分开来 [58]；接下来一个重要方向是根据观察到的外部状态来评估恢复，而不只评估对话中的修复。

近期的系统展示了这一挑战的不同侧面。MIRA 限制已提交执行的机器人运动量，同时确保语音仍可被打断；Yuma 的产品文档描述了客服通话进行期间发生的后端变更，以及随后给出的语音确认；CycleSpeech 则验证生成的语音是否符合预期的声音属性 [140], [152],[171]。最后一项关注的是语音保真度，而非外部操作；产品文档也不等同于独立的结果评估。与此同时，多模态模型发布以及 Snapdragon Gen 6 等端侧平台，将更多计算任务交由共享模型和本地硬件处理 [68], [90]。这些示例并未建立一条经过独立评估的追踪链，证明系统能从重叠语音出发，依次完成授权、执行、外部状态验证、恢复，最终获得经人工确认的结果。建立这条追踪链，是行动闭环面临的核心实验挑战。

表 IV 未解决的语音代理接口研究议程。每个方向都将相关工作与一个开放的系统问题、一项说明性实验和拟议的评估指标相结合。

研究方向	核心系统问题	说明性实验	关键评估指标
可修订的口语证据 [56], [85]	部分语音如何引导规划，同时保持依赖状态可修订？	在不同阶段引入自然纠正暂定规划的各个阶段；追踪其在信念、计划和待执行操作中的传播过程。	修订准确率；传播延迟；陈旧状态操作率。
身份、权限，以及持久状态 [18], [73]	身份和权限应如何与来源及有效期一同附加到持久决策状态？	跨会话追踪用户和记忆对象；引入过期权限、过期记录、纠正和删除请求；测试跨用户隔离。	更正与删除覆盖率；未经授权的复用和跨用户泄漏率。
闭合行动闭环 [58], [63]	规划器、执行器应如何以及验证器如何协调动作、在……情况下进行验证和恢复故障还是并发变更？	注入拒绝、超时、部分写入以及并发更新；追踪授权、拒答、执行、独立状态检查，以及通过取消、补偿或升级处置实现恢复。	经核实的任务成功；未经授权的行动率；恢复时间；残余副作用。
纵向协调 [201], [208]	记忆、恢复和交还方式能否适应用户需求，并随着会话增多而调整依赖程度？	针对不同群体开展重复会话研究，涵盖无障碍需求；改变记忆连续性、恢复支持和交还策略；追踪用户干预和任务恢复情况。	依赖校准；干预成本；任务恢复成功；可及性差距。

D. 纵向人机协同

随着语音智能体保留信息并执行任务，用户必须决定何时信任它们、何时介入。这个决定不仅取决于语音的自然度，也取决于智能体的时序、透明度、可及性以及错误处理方式。已有的小规模人类研究报告了适切性、可用性和接受度方面的差异 [81]，[201], [208]；口音和语言差异也会影响不同人群能否可靠地使用该接口 [99], [101]。

有关较长时期和真实场景的证据正逐渐涌现，尽管这些研究探讨的是不同问题。Cog-TiPRO 对十五名老年人在家中使用语音助手指令的情况进行了为期十八个月的每日追踪，以研究认知变化 [218]。这项研究确立了对语音助手使用情况的纵向观测，但并未验证智能体持久记忆的可靠性。一项关于招聘的自然现场实验还报告了由语音智能体进行面试所带来的后续就业结果，包括更高的录用率、更多的入职人数，以及员工留任率提高，同时测量生产率并未下降 [219]。这为真实世界中的使用情况及其对人类的影响提供了证据，但其本身并未检验跨会话记忆纠正或经独立验证的行动恢复。

长期协调还引出了更多问题。反复成功可能促使人们恰当依赖，也可能导致过度委派；反复失败则可能使人们变得谨慎，或选择放弃。个性化可以减少摩擦，但也可能让过时的假设在不知不觉中持续存在。即使偏好评分来自真实用户，也很难仅凭单次会话中的评分推断这些影响。

未来的评估可以跟踪同一批用户，同时记录智能体相关的状态转换。经过纠正的偏好在后续会话中是否仍保持纠正状态？一次不成功的动作是否会改变后续行为？当证据不足时，智能体是否会寻求澄清，用户能否迅速重新获得控制权？此类测试将人类行为与记忆、决策和恢复机制联系起来，而不会将类人性等同于质量。

E. 展望

这些方向共同的要求是可控耦合。系统可以在一轮交互结束前进行推理，但前提是暂定解释不会悄然变成不可逆的决策。只要持久记忆的来源和修订历史仍可供后续决策查阅，它就会发挥作用。当授权和验证能够与调用生成区分开来时，工具使用才会值得信赖；而适应必须让用户能够重新引导系统或重新获得控制权。这些是界面要求，而不是对某种特定架构的规定。

单一基准测试很可能无法涵盖所有这些方面。组件级评估对于隔离时序、识别、检索和工具使用机制仍然具有重要价值。还可以通过在选定边界之间保留因果轨迹的实验来加以补充：例如，在计划制定完成后才到达的纠正信息、受过时记忆影响的动作，或在物理操作之后进行观测、验证和交还。这样，无需每项研究都复现完整部署，就能对能动性闭环进行测量。

其目的并不是将听觉能动性归结为单一的成熟度评分，而是确定系统能够可靠处理哪些转换，以及哪些证据支持其成功执行动作的主张。

八、结论

语音智能体日益将实时语音、推理、记忆、工具使用、具身控制和以人为中心的保障措施结合起来。一旦口语证据能够改变持久状态或外部状态，仅罗列组件能力就不再足以描述重要的行为。听觉智能关注的是参与语音交互；听觉能动性 则始于这种交互与重大决策建立因果耦合之时。

两个接口使这种耦合变得具体。承诺边界将可修订证据与下游组件可能依赖的状态或动作区分开来。能动性闭环将经授权的行动与观测、验证、

恢复及其结果的交流联系起来。这些概念与架构无关；即便是统一模型，也必须能够回答：由谁或什么来负责决策，哪些证据为决策提供了依据，实际改变了什么状态，以及如何处理预期结果与观测结果之间的分歧。

评估应遵循同样的规范。组件指标和任务指标仍不可或缺，但它们的证据范围不应仅仅因为系统整合了更多功能就随之扩大。关于产生实际后果的能动性的主张，需要有能够追溯到后果的证据；部署主张则需要有能够触及用户及相关时间跨度的证据。进展将取决于能够保留从交互到结果之间因果轨迹的架构和评估，从而使系统的动作——而不仅仅是其语音或推理——能够得到审查并被追究责任。

参考文献

[1] A. D´efossez, L. Mazar´e, M. Orsini, A. Royer, P. P´erez, H. J´egou, E. Grave, and N. Zeghidour, 《Moshi：用于实时对话的语音文本基础模型》，2024年。 [在线]。可访问：https://arxiv.org/abs/2410.00037 。

[2] Z. Xie, J. Lang, Z. An, Y. Zhao, D. Yang, K. Li, Z. Ma, M. Lin, C. Miao, and S. Yan, 《Voicemem：用于实时交互的流式双脑记忆》，2026 [在线]。可访问：https://arxiv.org/abs/2608.26005 。

[3] L. M. Maben, G. G. Lakshmy, S. Radhakrishnan, S. Arora, and S. Watanabe, 《AURA：用于理解、推理和自动化工具使用的语音驱动任务智能体》，发表于*2025年IEEE自动语音识别与理解研讨会 (ASRU)*，2025年，第1–4页。 [在线]。可访问：https://arxiv.org/abs/2506.23049 。

[4] S. Wang, W. Yu, X. Chen, X. Tian, J. Zhang, L. Lu, and C. Zhang, 《端到端聆听、观察、说话与行动》，2025年。 [在线]。可访问：https://arxiv.org/abs/2510.16756 。

[5] S. Wang, Z. Jin, C. Tang, Q. Li, B. Li, C. Chen, Y. Hu, W. Yu, Y. Li, J. Zhuang 等，《迈向机器聆听与说话的通用听觉智能》，发表于《自然·机器智能》，第1–20页，2026年。[在线]。可访问：https://www.nature.com/articles/s42256-026-01281 -1

[6] K. Hu, E. Hosseini-Asl, C. Chen, E. Casanova, S. Ghosh, P. ˙Zelasko, Z. Chen, J. Li, J. Balam, and B. Ginsburg, 《面向语音到语音语言模型的高效直接双工建模》，发表于*Interspeech 2025*，2025年，第2715–2719页。 [在线]。可访问：https://arxiv.org/abs/2505.15670 。

[7] X. Wang, Y. Li, C. Fu, Y. Shen, L. Xie, K. Li, X. Sun, and L. Ma, 《Freeze-omni：采用冻结大语言模型的智能低延迟语音到语音对话模型》，2024年。 [在线]。可访问：https://arxiv.org/abs/2411.00774 。

[8] R. B. Sinha, 《AnovaX：具备大语言模型规划、类型化执行器和自适应恢复能力的本地多智能体语音助手》，2026年。 [在线]。可访问：https://arxiv.org/abs/2607.15367 。

[9] Z. Ma, B. VanDerPloeg, C.-P. Bara, H. Yidong, E.-I. Kim, F. Gervits, M. Marge, and J. Chai, 《DOROTHIE：交互式自动驾驶智能体应对意外情况的口语对话》，发表于计算语言学协会论文集：*EMNLP 2022*，2022年，第4800–4822页。 [在线]。可访问：https://arxiv.org/abs/2210.12511 。

[10] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. de Chaumont Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi, A. Tudor, M. Velimirovi´c, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, 《AudioPaLM：能够说话和聆听的大语言模型》，2023年。 [在线]。可访问：https://arxiv.org/abs/2306.12925

[11] W. Yu, S. Wang, X. Yang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, Y. Wang, and C. Zhang, 《SALMONN-omni：用于全双工语音理解与生成的无编解码器大语言模型》，2024年。 [在线]。可访问：https://arxiv.org/abs/2411.18138

[12] Q. Chen, Y. Chen, Y. Chen, M. Chen, Y. Chen, C. Deng, Z. Du, R. Gao, C. Gao, Z. Gao, Y. Li, X. Lv, J. Liu, H. Luo, B. Ma, C. Ni, X. Shi, J. Tang, H. Wang, H. Wang, W. Wang, Y. Wang, Y. Xu, F. Yu, Z. Yan, Y. Yang, B. Yang, X. Yang, G. Yang, T. Zhao, Q. Zhang, S. Zhang, N. Zhao, P. Zhang, C. Zhang, and J. Zhou, 《MinMo：用于无缝语音交互的多模态大语言模型》，2025年。 [在线]。可访问：https://arxiv.org/abs/2501.06282

[13] Y. Cao, N. Potdar, and A. R. Avila, 《序列端到端意图与槽位标签分类及定位》，arXiv预印本 arXiv:2106.04660，2021年。

[14] W. Wu, W. Guan, K. Wang, P. Chen, Z. Zha, J. Li, J. Fang, L. Li, and Q. Hong, 《Phoenix-VAD：用于全双工语音交互的流式语义端点检测》，2025年。 [在线]。可访问：https://arxiv.org/abs/2509.20410 。

[15] S. Arora, J. Tian, H. Futami, J. weon Jung, J. Shi, Y. Kashiwagi, E. Tsunoo, and S. Watanabe, 《面向开放式端到端口语对话系统的思维链训练》，2025年。 [在线]。可访问：https://arxiv.org/abs/2506.00722 。

[16] W. Zou, Y. Miao, Z. Ma, J. Xu, J. Gao, J. Hao, R. He, and J. Xu, 《Lts-voiceagent：通过语义触发与增量推理实现高效流式语音交互的聆听-思考-说话框架》，2026 [在线]。可访问：https://arxiv.org/abs/2601.19952 。

[17] J. Kirmayr, L. Stappen, P. Schneider, F. Matthes, and E. Andr´e, 《Carmem：通过类别界定增强大语言模型语音助手的长期记忆》，2025年。 [在线]。可访问：https://arxiv.org/abs/2501.09645 。

[18] M. Al-Ratrout, P. U. Ravva, S. Sharmin, A. Raikwar, J. Y. Shin, and R. L. Barmaki, 《Afa：用于防止多用户对话中人设混淆的身份感知记忆》，arXiv预印本 arXiv:2604.25022，2026年。

[19] R. Pahwa, A. Beedu, P. Priye, R. Gandhi, S. Takawale, A. Baijal, and Z. Yang, 《Audio2tool：开口说、调用、行动——用于语音工具使用基准评测的数据集》，2026年。 [在线]。可访问：https://arxiv.org/abs/2604.22821 。

[20] H. Zhang, J. Chen, D. Wu, Y. Li, Y. Zhang, X. T. Zhang, C. Liu, Q. Lin, Y. Peng, H. Liu, E. S. Chng, C. Yan, B. Wu, Y. Huang, X. Yang, and F. Tian, 《DuplexSLA：具备同步语音、语言与动作的全双工口语语言模型》，2026年。 [在线]。可访问：https://arxiv.org/abs/2605.20755 。

[21] S. Wang, J. Fu, F. Liu, X. He, H. Wu, J. Shi, K. Huang, Z. Fei, J. Gong, Z. Wu, Y.-G. Jiang, S.-K. Ng, T.-S. Chua, and X. Qiu, 《Roboomni：在全模态上下文中主动进行机器人操作》，2025年。 [在线]。可访问：https://arxiv.org/abs/2510.23763

[22] S. Yang, Y. Huang, W. Cai, S. Sun, F. Fang, Y. He, Y. Xie, J. Deng, H. Zhang, J. Song, and Z. Zhang, 《第一人称视角副驾驶：用于辅助型第一人称视角人工智能的原生网页智能眼镜智能体》，2026年。 [在线]。可访问：https://arxiv.org/abs/2603.01104 。

[23] A. Papangelis, M. Namazifar, C. Khatri, Y.-C. Wang, P. Molino, and G. Tur, 《Plato 对话系统：灵活的对话式人工智能研究平台》，arXiv预印本 arXiv:2001.06463，2020年。

[24] V. Pal, F. Guillot, M. Shrivastava, J.-M. Renders, and L. Besacier, 《利用词混淆网络对话状态跟踪中的自动语音识别歧义建模》，发表于 Interspeech 2020，2020年，第1545–1549页。 [在线]。可访问：https://arxiv.org/abs/2002.00768

[25] J. Chen, R. Zhang, Y. Mao, and J. Xu, 《利用时间表达网络进行神经对话状态跟踪》，发表于计算语言学协会论文集：EMNLP 2020，2020年，第1570–1579页。 [在线]。可访问：https://arxiv.org/abs/2009.07615 。

[26] H. H. Clark and S. E. Brennan, 《交流中的共同基础》。1991年。

[27] P. Dillenbourg, D. Traum, and D. Schneider, 《多模态任务导向协作中的共同基础》，发表于欧洲人工智能教育会议论文集，1996年，第401–407页。

[28] S. Young, M. Ga˘si´c, B. Thomson, and J. D. Williams, 《基于部分可观测马尔可夫决策过程的统计口语对话系统：综述》，《IEEE会议录》，第101卷，第5期，第1160–1179页，2013年。

[29] Z. Chen, L. Chen, X. Zhou, and K. Yu, 《用于在线对话状态跟踪的深度强化学习》，发表于《人机语音通信》（《计算机与信息科学通信》）。Springer Nature Singapore，2023年，第278–292页。 [在线]。可访问：https://arxiv.org/abs/2009.10321 。

[30] Z. Chen, L. Chen, X. Liu, and K. Yu, 《面向通用对话管理的分布式结构化演员-评论家强化学习》，《IEEE/ACM 音频、语音与语言处理汇刊》，第28卷，第2400–2411, 2020 [在线]。可访问：https://arxiv.org/abs/2009.10326 –页。在线。

[31] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, 《面向语音智能体的对话式端到端文本转语音

," 于*2021*年 *IEEE* 口语

技术研讨会（SLT）。IEEE，2021年，第 403–409 页。 [在线]。获取地址：https://arxiv.org/abs/2005 10438 。Y. Jia、M. T. Ramanovich、T. Remez 和 R. Pomerantz，“Translatotron 2：高质量的直接语音到语音翻译与语音保留”，2021 [ ] 2107 08661 。在线。获取地址：https://arxiv.org/abs/。A. Lee、P.-J. Chen、C. Wang、J. Gu、S. Popuri、X. Ma、A. Polyak、Y. Adi、Q. He、Y. Tang 等，“基于离散单元的直接语音到语音翻译”，载于计算语言学协会第 60 届年会论文集（第 1 卷：长篇论文），2022年，第 3327 3339 [ ] 2107 05604 – 页。在线。获取地址：https://arxiv.org/abs/。H. Inaguma、S. Popuri、I. Kulikov、P.-J. Chen、C. Wang、Y.-A. Chung、Y. Tang、A. Lee、S. Watanabe 和 J. Pino，“Unity：基于离散单元的两阶段直接语音到语音翻译”，2022年。 [在线]。获取地址：https://arxiv.org/abs/2212 08055 。E. Nachmani、A. Levkovitch、Y. Ding、C. Asawaroengchai、H. Zen 和 M. T. Ramanovich，“Translatotron 3：使用单语数据进行语音到语音翻译”，2023年。 [在线]。获取地址：https://arxiv.org/abs/2305 17547 。S. Communication、L. Barrault、Y.-A. Chung、M. C. Meglioli、D. Dale、N. Dong、P.-A. Duquenne、H. Elsahar、H. Gong、K. Heffernan、J. Hoffman、C. Klaiber、P. Li、D. Licht、J. Maillard、A. Rakotoarison、K. R. Sadagopan、G. Wenzek、E. Ye、B. Akula、P.-J. Chen、N. E. Hachem、B. Ellis、G. M. Gonzalez、J. Haaheim、P. Hansanti、R. Howes、B. Huang、M.-J. Hwang、H. Inaguma、S. Jain、E. Kalbassi、A. Kallet、I. Kulikov、J. Lam、D. Li、X. Ma、R. Mavlyutov、B. Peloquin、M. Ramadan、A. Ramakrishnan、A. Sun、K. Tran、T. Tran、I. Tufanov、V. Vogeti、C. Wood、Y. Yang、B. Yu、P. Andrews、C. Balioglu、M. R. Costa-juss`a、O. Celebi、M. Elbayad、C. Gao、F. Guzm´an、J. Kao、A. Lee、A. Mourachko、J. Pino、S. Popuri、C. Ropers、S. Saleem、H. Schwenk、P. Tomasello、C. Wang、J. Wang 和 S. Wang，“Seamlessm4t：大规模多语言与多模态机器翻译”，2023年。 [ ] 2308 11596 在线。获取地址：https://arxiv.org/abs/。D. Zhang、S. Li、X. Zhang、J. Zhan、P. Wang、Y. Zhou 和 X. Qiu，“Speechgpt：赋能大语言模型具备内在的跨模态对话能力”，载于计算语言学协会论文集：EMNLP 2023，2023年，第 15 757–15 773 页。[38] T. A. Nguyen、B. Muller、B. Yu、M. R. Costa-Jussa、M. Elbayad、S. Popuri、C. Ropers、P.-A. Duquenne、R. Algayres、R. Mavlyutov 等，“Spirit-lm：交错式口语与书面语语言模型”，计算语言学协会会刊，第 13 卷，52，2025年，30页。–。[39] T. A. Nguyen、E. Kharitonov、J. Copet、Y. Adi、W.-N. Hsu、A. Elkahky、P. Tomasello、R. Algayres、B. Sagot、A. Mohamed 和 E. Dupoux，“生成式口语对话语言建模”，2022年。 [在线]。获取地址：https://arxiv.org/abs/2203 16502 。D. Schlangen 和 G. Skantze，“增量式对话处理的一般抽象模型”，Dialogue & Discourse，第 2 卷，第 83–111 页，2011年。[41] G. Skantze 和 A. Hjalmarsson，“面向对话系统中的增量式语音生成”，载于 SIGDIAL 2010 会议论文集，2010年，第 1–8 页。[42] M. Roddy 和 N. Harte，“对话响应时机的神经生成”，载于计算语言学协会第 58 届年会论文集，2020年，第 2442–2452 页。[在线]。获取地址：https://aclanthology.org/2020.acl-main.221/ [43] S. Li、A. Paranjape 和 C. D. Manning，“我什么时候可以说话？预测口语对话代理的发言起始点”，2022年。 [在线]。获取地址：https://arxiv.org/abs/2208 03812 。G. Skantze 和 B. Irfan，“将通用轮流发言模型应用于对话式人机交互”，载于 2025年第 20 届 ACM/IEEE 人机交互国际会议（HRI）。IEEE，2025年，第 859–868 页。[45] H. J. Asl、A. Edraki、M. G. Nejad、M. Asgharian、M. Sadeghi、Y. Yu 和 V. P. Nia，“Hywa：面向全双工语音助手的保留架构的个性化语音活动检测”，2026年。 [在线]。获取地址：https://arxiv.org/abs/2510 12947 。G. Li、C. Wang、H. Xue、S. Wang、D. Gao、Z. Zhang、Y. Lin、W. Li、L. Xiao、Z. Fu 和 L. Xie，“Easy turn：融合声学与语言模态，实现全双工口语对话系统中的稳健轮流发言”，载于 ICASSP 2026 - 2026年 IEEE声学、语音与信号处理国际会议（ICASSP），2026年，第 16 957 16 961 [ ] 2509 23938 – 页。在线。获取地址：https://arxiv.org/abs/。D. Bekal、S. Srinivasan、S. Bodapati、S. Ronanki 和 K. Kirchhoff，“利用上下文线索实现口语对话系统的设备指向性”，2022 [ ] 2211 13280 。在线。获取地址：https://arxiv.org/abs/。A. Zeng、Z. Du、M. Liu、K. Wang、S. Jiang、L. Zhao、Y. Dong 和 J. Tang，“GLM-4-Voice：迈向智能且类人端到端口语聊天机器人”，2024年。 [在线]。可访问：https://arxiv.org/abs/2412 02612 。W. Wang、C. Li、L. Zhang、Y. Zhao、Y. Zou、H. Li、M. Cui、H. Zhang、K. Wei、L. Xu、Z. Huang、J. Xu、J. Hu、X. He、Z. Xie、J. Kang、Y. Chen、M. Yu、D. Yu、R. Chen、L. Di、S. Feng、N. Hu、Y. Liu、B. Wang 和 S. Yang，“Covo-audio 技术报告”，2026年。 [在线]。可访问：https://arxiv.org/abs/2602 09823 。Q. Fang、Y. Zhou、S. Guo、S. Zhang 和 Y. Feng，“LLaMA-Omni 2：基于大语言模型、采用自回归流式语音合成的实时口语聊天机器人”，载于《计算语言学协会第63届年会论文集》（第1卷：长篇论文），2025年，第18 617–18 629页。[51] S. Liu、Y. Lu、W. Fang、J. Wang 和 Z. Xie，“Sync-llm：使用分层语言模型生成大规模合成电路代码”，载于《2025年自然语言处理实证方法会议论文集》，2025年，第17 361–17 376页。[52] Q. Zhang、L. Cheng、C. Deng、Q. Chen、W. Wang、S. Zheng、J. Liu、H. Yu、C.-H. Tan、Z. Du 等，“OmniFlatten：用于无缝语音对话的端到端 GPT 模型”，载于《计算语言学协会第63届年会论文集》（第1卷：长篇论文），2025年，第14 570–14 580页。[53] B. Liao、Y. Xu、J. Ou、K. Yang、W. Jian、P. Wan 和 D. Zhang，“FlexDuo：一种在语音对话系统中实现全双工能力的可插拔系统”，2025年。 [在线]。可访问：https://arxiv.org/abs/2502 13472 。G.-T. Lin、S.-Y. S. Kuan、Q. Wang、J. Lian、T. Li、S. Watanabe 和 H. yi Lee，“Full-duplex-bench v1.5：评估全双工语音模型的重叠处理能力”，载于《ICASSP 2026——2026年电气与电子工程师协会声学、语音与信号处理国际会议论文集》（ICASSP），2026年，第 19 447 19 451 [ ] 2507 23159 –页。在线。可访问：https://arxiv.org/abs/。G.-T. Lin、S.-Y. S. Kuan、J. Shi、K.-W. Chang、S. Arora、S. Watanabe 和 H. yi Lee，“Full-duplex-bench-v2：使用自动评估器对双工对话系统进行多轮评估的框架”，载于《计算语言学协会第64届年会论文集》（第2卷：短篇论文），2026年，第27–36页。 [ ] 2510 07838 在线。可访问：https://arxiv.org/abs/。G.-T. Lin、C. Chen、Z. Chen 和 H. yi Lee，“Full-duplex-bench-v3：在真实场景中的言语不流畅条件下对全双工语音智能体的工具使用进行基准评测”，2026 [ ] 2604 04847 。在线。可访问：https://arxiv.org/abs/。Y. Peng、Y.-W. Chao、D. Ng、Y. Ma、C. Ni、B. Ma 和 E. S. Chng，“FD-Bench：为全双工口语对话系统设计的全双工基准测试流程”，载于 Interspeech 2025，2025年，第 176 180 [ ] 2507 19040 –页。在线。可访问：https://arxiv.org/abs/。A. Salimi、W. Ma、Y. Tang、D. Shen、M. Li 和 A. Smola，“IHBench：使用结构化工作流评估语音智能体的中断后恢复能力”，2026 [ ] 2606 19595 。在线。可访问：https://arxiv.org/abs/。S. Yao、J. Zhao、D. Yu、N. Du、I. Shafran、K. Narasimhan 和 Y. Cao，“ReAct：在语言模型中协同推理与行动”，arXiv 预印本 arXiv:2210.03629，2022年。[60] R. Huang、M. Li、D. Yang、J. Shi、X. Chang、Z. Ye、Y. Wu、Z. Hong、J. Huang、J. Liu、Y. Ren、Z. Zhao 和 S. Watanabe，“AudioGPT：理解和生成语音、音乐、声音及会说话的人头视频”，2023 [ ] 2304 12995 。在线。可访问：https://arxiv.org/。S. Arora、J. Tian、H. Futami、J. Shi、Y. Kashiwagi、E. Tsunoo 和 S. Watanabe，“流式全双工端到端语音对话系统中的思维链推理”，2025年。 [在线]。可访问：https://arxiv.org/abs/2510 02066 。M. T. R. Laskar、X.-Y. Fu、S. S. Sarfjoo、Q. McNamara、J. Robertson 和 S. B. TN，“从文本到语音：用于评估工具调用大语言模型智能体的可复现、可验证框架”，2026年。 [在线]。可访问：https://arxiv.org/abs/2605 15104 。S. Ray、K. Dhandhania、V. Barres 和 K. Narasimhan，“τ-Voice：在现实世界领域对全双工语音智能体进行基准评测”，2026年。 [ ] 2603 13686 在线。可访问：https://arxiv.org/abs/。A. Dalmia、S. Midha 和 J. Doshi，“Formbharo：为印度农村地区设计并评估用于对话式表单填写的语音智能体”，2026 [ ] 2608 06027 。在线。可访问：https://arxiv.org/。S. P. Cherumanal、L. Tian、F. M. Abushaqra、A. F. M. de Paula、K. Ji、D. Hettiachchi、J. R. Trippas、H. Ali、F. Scholer 和 D. Spina，“Walert：通过构建和评估大语言模型驱动的聊天机器人，将对话式搜索知识付诸行动”，2024年。 [在线]。可访问：https://arxiv.org/abs/2401 07216 。S. Zhong、E. Gatti、J. Hardwick、M. Ribul、Y. Cho 和 M. Obrist，“大语言模型介导的领域专用语音智能体：以 textilebot 为例”，2024年。 [在线]。可访问：https://arxiv.org/abs/2406.10590

[67] S. Yao、N. Shinn、P. Razavi 和 K. Narasimhan，“tau-bench：现实世界领域中工具-智能体-用户交互的基准测试”，arXiv 预印本 arXiv:2406.12045，2024年。[68] 小米 MiMo 团队，“MiMo-V2.5 系列发布”，小米 MiMo 文档，2026年4月，官方模型发布文档；访问于2026年9月。 [在线]。可访问：https://mimo.mi.com/do cs/en-US/updates/model [69] Q. Team，“Qwen3.5-omni 技术报告”，arXiv 预印本 arXiv:2604.15804，2026年。[70] Y. Chen、S. Ji、H. Wang、Z. Wang、S. Chen、J. He、J. Xu 和 Z. Zhao，“Wavrag：面向口语对话模型的音频集成检索增强生成”，载于《计算语言学协会第63届年会论文集》（第1卷：长篇论文），2025年，第12 505–12 523页。 [在线]。可访问：https://arxiv.org/abs/ 2502 14727 。Z. Rackauckas 和 J. Hirschberg[71]，“Voxrag：迈向口语问答中的无转录文本 RAG 系统”，2025年。 [在线]。可访问：https://arxiv.org/abs/ 2505 17326 。S. Kuroki、Y. Kubo、T. Akiba 和 Y. Tang[72]，“Kame：用于增强实时语音到语音对话式人工智能知识的串联架构”，载于《ICASSP 2026——2026年电气与电子工程师协会声学、语音与信号处理国际会议论文集》（ICASSP）。电气与电子工程师协会，2026年，第19 362–19 366页。 [在线]。可访问：https://arxiv.org/abs/2510.02327[73] A. Gosai、T. Vuong、U. Tyagi、S. Li、W. You、M. Bavare、A. Uc.ar、Z. Fang、B. Jang、B. Liu 和 Y. He，“Audio multichallenge：在自然人类交互中对语音对话系统进行多轮评估”，载于《计算语言学协会第64届年会论文集》（第1卷：长篇论文），2026年，第35 740–35 770 [ ] 2512 14865 –页。在线。可访问：https://arxiv.org/abs/。D. Wu、H. Wang、W. Yu、Y. Zhang、K.-W. Chang 和 D. Yu[74]，“LongMemEval：对聊天助手的长期交互记忆进行基准评测”，arXiv 预印本 arXiv:2410.10813，2024年。[75] M. Chen、K. Wang、L. Lu、J. Zhang 和 T. Zhang，“通过与上下文无关且不可感知的听觉提示注入劫持大型音频语言模型”，载于《2026年电气与电子工程师协会安全与隐私研讨会》（SP），2026年，第310–328页。 [在线]。可访问：https: //arxiv.org/abs/2604.14604[76] A. Djanibekov、N. Mukhituly、K. Inui、H. Aldarmaki 和 N. Lukas，“Spirit：通过修补语音语言模型抵御越狱攻击”，载于《2025年自然语言处理实证方法会议论文集》，2025年，第14 503–14 520页。 [在线]。可访问：https://arxiv.org/abs/ 2505 13541 。Y. Wang、H. Liu、D. Chen、X. Zhang 和 Z. Wu[77]，“Voxprivacy：评估语音语言模型交互隐私的基准测试”，2026 [ ] 2601 19956 。在线。可访问：https://arxiv.org/abs/。X. Li、P.-Y. Chen 和 W. Wei[78]，“Aegis：迈向人工智能语音智能体的治理、完整性与安全”，2026年。 [在线]。可访问：https://arxiv.org/abs/2602.07379[79] M. Ahn、A. Brohan、N. Brown、Y. Chebotar、O. Cortes、B. David、C. Finn、C. Fu、K. Gopalakrishnan、K. Hausman 等，“照我能做的去做，而不是照我说的去做：在机器人可供性中建立语言基础”，arXiv 预印本 arXiv:2204.01691，2022年。[80] J. Acosta、S. Adams、J. Kernbach、R. Hardy、S. E. Kim、L. Luo、X. Zhang、S. Johri、M. Baharoon 和 P. Rajpurkar，“用于临床评估的语音引导编排智能（VOICE）：用于院前卒中评估的语音 AI 智能体系统”，2025年。 [在线]。可访问：https://arxiv.org/abs/ 2507 22898 。X. Xu、B. Yao、Z. Yang、S. Zhang[81]、E. Rogers、S. Intille、N. Shara、G. Gao 和 D. Wang，“Talk2care：利用大语言模型促进异步患者-医疗服务提供者交流”，《AAAI 研讨会系列》，第4卷，第1期，第146–151页，2024年。 [在线]。可访问：https://arxiv.org/abs/ 2309 09357 。Q. Fang、S. Guo 和 Y. Feng[82]，“BayLing-Duplex：使用单个自回归大语言模型实现原生全双工语音对话”，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2606 14528 。J. Yang、Y. Fujita 和 Y. Sudo[83]，“DuplexCascade：采用无 VAD 级联 ASR-LLM-TTS 流水线和微轮次优化的全双工语音到语音对话”，2026年。 [在线]。可访问：https: //arxiv.org/abs/ 2603 09180 。J. Yang、P. Wang、Y. Zhu、M. Feng、M. Chen 和 X. He[84]，“用于人机对话中轮流发言预测的门控多模态融合与对比学习”，载于《ICASSP 2022——2022年电气与电子工程师协会声学、语音与信号处理国际会议论文集》（ICASSP）。电气与电子工程师协会，2022年，第7747–7751页。 [在线]。可访问：https://arxiv.org/abs/2204.10172 [85] S. Udupa、S. Watanabe、P. Schwarz 和 J. Cernocky，《低延迟口语对话的端点预判》，2026年。 [在线]。获取地址：https://arxiv.org/abs/2606.13450[86] ——，《使用神经音频编解码器和标签延迟训练的口语对话流式端点检测器》，2025年。 [在线]。获取地址：https://arxiv.org/abs/ 2506 07081 。W. Cui、L. Zhu、X. Li、Z. Guo、H. Bai、L. Hou[87] 和 I. King，《Turnguide：通过动态轮次级文本-语音交错，提升富有意义的全双工语音交互》，2025年。 [在线]。获取地址：https://arxiv.org/abs/ 2508 07375 。Y. Wang、T.-H. Zhang、M. Cai、Y. Ren、Z. Jiang[88]、X. Wang、Z. Wang、P. Zhou、K. Zhan 和 X. Qian，《聆听、调用并理解：面向大型音频语言模型的技能调用多模态智能体》，2026年。 [ ] 2608 01881 在线。获取地址：https://arxiv.org/abs/。E. Studerus、V. J. Zhong 和 S. Vonschallen[89]，《面向 Pepper 机器人的低延迟大语言模型驱动多模态交互框架》，2026 [ ] 2603 21013 。在线。获取地址：https://arxiv.org/abs/。Qualcomm Technologies, Inc.[90]，《Snapdragon凭借两款全球最快的移动系统级芯片，引领智能体人工智能时代，为新一代智能手机提供强劲动力》，高通新闻稿，2026年9月，2026年9月22日。[在线]。获取地址：https://www.qualcomm.com/news/releases/2026/09 / snapdragon-leads-the-agentic-ai-age-with-two-of-the-world-s-fast[91] ——，《Snapdragon 8 Elite Extreme Gen 6移动平台》，官方产品文档，2026年，查阅于2026年9月。 [在线]。获取地址：https://www.qualcomm.com/smartphones/products/8-series/ snapdragon-8-elite-extreme-gen-6-mobile-platform [92] I. Mehta，《高通推出两款全新智能手机芯片，主打人工智能》，TechCrunch，2026年9月，2026年9月22日。[在线]。获取地址：https://techcrunch.com/2026/09/22/qualcomm-launches-two - new-smartphone-chips-with-emphasis-on-ai/ [93] S. Watanabe、T. Hori、S. Karita、T. Hayashi、J. Nishitoba、Y. Unno、N. E. Y. Soplin、J. Heymann、M. Wiesner、N. Chen 等，《ESPnet：端到端语音处理工具包》，arXiv预印本 arXiv:1804 00015，2018年。。[94] S. E. Eskimez、T. Yoshioka、A. Ju、M. Tang、T. Parnamaa 和 H. Wang，《实时联合个性化语音增强与声学回声消除》，2022年。 [在线]。获取地址：https: //arxiv.org/abs/ 2211 02773 。C. Liu、S. Xue、H. Yan、X. Liang 和 Z. Xue[95]，《LMPAN：用于全双工声学回声消除与噪声抑制的轻量级多路径对齐网络》，2026年。 [在线]。获取地址：https://arxiv.org/abs/ 2607 02062 。A. Gulati、J. Qin、C.-C. Chiu、N. Parmar[96]、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 等，《Conformer：用于语音识别的卷积增强型Transformer》，arXiv预印本 arXiv:2005 08100，2020年。。[97] A. Radford、J. W. Kim、T. Xu、G. Brockman、C. McLeavey 和 I. Sutskever，《通过大规模弱监督实现鲁棒语音识别》，载于《国际机器学习会议》。PMLR，2023年，第28 492–28 518页。[98] H. Liu、Y. Wang、Z. Cheng、R. Wu、Q. Gu、Y. Wang 和 Y. Wang，《Cs3-bench：评估并增强适用于中英语码转换的语音到语音大语言模型》，2025年。 [在线]。获取地址：https://arxiv.org/abs/ 2510 07881 。S. Baroudi、Y. Labrak、S. Kumar、J. Kalda、S. Burdisso[99]、P. Cyrta、J. I. Alvarez-Trejos、P. Motlicek、H. Bredin 和 R. Marxer，《医生还是患者？通过协同说话人分离与自动语音识别，提取语码转换印地英语中的医疗状况》，2026年。 [在线]。获取地址：https://arxiv.org/abs/ 2603 06373 。M. Feng、W. Tan 和 M. Pilanci[100]，《语音识别中的凸优化低资源口音鲁棒语言检测》，2026年。 [在线]。获取地址：https://arxiv.org/abs/ 2605 23235 。M. Maltais、Y. Jeon、M. Ma、S. H. Muhammad[101]、I. Abdulmumin、M. I. Mukhtar、D. Abolade、J. Okepefi、J. Sewedo 和 D. I. Adelani，《NaijaS2ST：面向低资源尼日利亚语言语音到语音翻译的多口音基准测试集》，2026年。 [在线]。获取地址：https://arxiv.org/abs/ 2604 16287 。D. Lanasri、R. I. A. Aouchiche、A. Remmide、F. Taki[102] 和 A. Kemmoum，《Dziri Voicebot：面向阿尔及利亚方言的端到端低资源语音到语音会话系统》，2026年。 [在线]。获取地址：https://arxiv.org/abs/ 2606 26003 。M. N. Ali、D. Falavigna[103] 和 A. Brutti，《SpeechLLM与联邦学习结合，实现端到端自动语音识别：英语和意大利语案例研究》，载于2026年第二届国际联邦学习与智能计算系统会议（FLICS）。电气与电子工程师协会，2026年，第146–153页。 [在线]。获取地址：https://arxiv.org/abs/2607.25716

[104] G. Chollet、H. Sansen、Y. Tevissen、J. Boudy、M. Hariz、C. Lohr 和 F. Yassa，《面向家庭及其他场景的隐私保护个人助理：具备端侧说话人分离功能的口语对话系统》，载于《人机交互与新兴技术》（IHIET-AI 2024），第 120, 2024 [ ] 2401 01146 卷。在线。获取地址：https://arxiv.org/abs/。[105] K. Inoue、D. Lala、M. Elmers、K. Ochi 和 T. Kawahara，《面向多模态多方对话中受话人识别的大语言模型基准测试》，2025 [ ] 2501 16643 。在线。获取地址：https://arxiv.org/abs/。[106] H. Bredin、R. Yin、J. M. Coria、G. Gelly、P. Korshunov、M. Lavechin、D. Fustes、H. Titeux、W. Bouaziz 和 M.-P. Gill，《Pyannote.audio：用于说话人分离的神经网络基础模块》，载于ICASSP 2020——2020年电气与电子工程师协会声学、语音与信号处理国际会议（ICASSP）。电气与电子工程师协会，2020年，第7124–7128页。[107] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur，《X-vector：用于说话人识别的鲁棒深度神经网络嵌入》，载于2018年电气与电子工程师协会声学、语音与信号处理国际会议（ICASSP）。电气与电子工程师协会，2018年，第5329–5333页。[108] G. P. Spithourakis、I. Vuli´c、M. Lis、I. Casanueva 和 P. Budzianowski，《Evi：用于基于知识的注册、验证和身份识别的多语言口语对话任务与数据集》，载于《计算语言学协会：NAACL 2022论文集》，2022年，第 1647 1659 [ ] 2204 13496 –页。在线。获取地址：https://arxiv.org/abs/。[109] E. Ekstedt 和 G. Skantze，《增量式口语对话系统中的轮次完成预判》，载于《第22届话语与对话特别兴趣小组年会论文集》，2021年，第431–437页。[110] A. Raux 和 M. Eskenazi，《优化面向任务的口语对话系统的轮流发言行为》，《ACM语音与语言处理汇刊》（TSLP），第9卷，第1期，第1–23页，2012年。[111] G. Skantze，《会话系统与人机交互中的轮流发言：综述》，《Computer Speech & Language》，第67卷，101178，2021年，第页。[112] H. Khouzaimi、R. Laroche 和 F. Lef`evre，《用于增量式口语对话系统轮流发言管理的强化学习》，载于IJCAI，2016年，第2831–2837页。[113] S. C. Levinson，《人类交流中的轮流发言——起源及其对语言处理的影响》，《Trends in Cognitive Sciences》，第20卷，第1期，第6–14页，2016年。[114] K. Inoue、D. Lala、G. Skantze 和 T. Kawahara，《嗯、呃、哦：通过微调语音活动预测实现连续实时反馈语预测》，2024年。 [在线]。获取地址：https://arxiv.org/abs/2410 15929 。J. Wang、L. Chen、A. Khare、A. Raju、P. Dheram、D. He、M. Wu、A. Stolcke 和 V. Ravichandran，《通过声学与大语言模型融合进行轮流发言和反馈语预测》，2024年。 [ ] 2401 14717 在线。获取地址：https://arxiv.org/abs/。[116] K. Inoue、M. Elmers、Y. Fu、Z. H. Pang、T. Mori、D. Lala、K. Ochi 和 T. Kawahara，《多语言连续反馈语预测：一项跨语言研究》，2025年。 [在线]。获取地址：https://arxiv.org/abs/2512 14085 。M. Heldner 和 J. Edlund，《对话中的停顿、间隔与重叠》，《Journal of Phonetics》，第38卷，第4期，第555–568页，2010年。[118] H. Khouzaimi、R. Laroche 和 F. Lefevre，《增量式对话系统中的轮流发言现象》，载于《2015年自然语言处理实证方法会议论文集》，2015年，第1890–1895页。[119] G. Yang、Y. Pan、S. Qiu 和 N. Bai，《Jal-turn：用于全双工口语对话系统实时鲁棒轮流发言检测的联合声学-语言建模》，2026年。 [在线]。获取地址：https://arxiv.org/abs/2603 26515 。K. Fu、R. Wen、A. Lin、S. Qin、R. Gan、H. Wang 和 Q. Wang，《X2-turn：用于联合流式自动语音识别与轮次状态预测的帧同步双头建模》，2026年。 [在线]。获取地址：https://arxiv.org/abs/2608 10878 。C. Wang、H. Xue、M. Shao、C. He、J. Hu、S. Wang、B. Wu、Y. Ji、J. Zheng、R. Chen、Z. Zhu 和 L. Xie，《Fastturn：融合声学与流式语义线索，实现低延迟鲁棒轮次检测》，2026 [ ] 2604 01897 。在线。获取地址：https://arxiv.org/abs/。[122] H. Shen、Y. Wang、S. Huang、W. Zou 和 Y. Chen，《CodeTT：用于轮流发言评估的上下文感知决策基准测试》，2026年。 [ ] 2603 25434 在线。获取地址：https://arxiv.org/abs/。[123] K. Xia、B. Mu、X. Shi、J. Xu 和 L. Xie，《口语对话系统中的语义感知打断检测：基准测试、指标与模型》，2026 [ ] 2603 24144 。在线。获取地址：https://arxiv.org/abs/。[124] Z. He、H.-N. Liang 和 C. Liang，《Tap-to-adapt：为语音智能体学习与用户对齐的响应时机》，2026年。 [在线]。获取地址：https://arxiv.org/abs/2603.14449

[125] J. Lee、J. Pyo、J. Park 和 Y. Jo，《Spokenus：面向任务导向对话的口语用户模拟器》，2026年。 [在线]。可访问：https://arxiv.org/abs/2603 16783 。T. Zhong、J. Deng、N. Kuzmin、Y. Zhu、T. Cao、T. Tsoi、Z. Tan、S. Lui 和 X. Liu，《Iraf：用于噪声鲁棒端到端全双工口语对话系统的抗干扰自适应融合》，2026年。 [ ] 2606 06559 在线。可访问：https://arxiv.org/abs/。R. Roy、J. Raiman、S. gil Lee、T.-D. Ene、R. Kirby、S. Kim、J. Kim 和 B. Catanzaro，《Personaplex：全双工对话语音模型的语音与角色控制》，2026年。 [在线]。可访问：https://arxiv.org/abs/2602 06053 。Y. Wang、R. Skerry-Ryan、D. Stanton、Y. Wu、R. J. Weiss、N. Jaitly、Z. Yang、Y. Xiao、Z. Chen、S. Bengio 等，《Tacotron：迈向端到端语音合成》，arXiv预印本 arXiv:1703.10135，2017年。[129] J. Shen、R. Pang、R. J. Weiss、M. Schuster、N. Jaitly、Z. Yang、Z. Chen、Y. Zhang、Y. Wang、R. Skerrv-Ryan 等，《通过以梅尔频谱图预测为条件的Wavenet 实现自然 TTS 合成》，载于《2018年IEEE声学、语音与信号处理国际会议（ICASSP）》。电气与电子工程师协会，2018年，第4779–4783页。[130] J. Kong、J. Kim 和 J. Bae，《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》，第33卷，第17022–17033页，2020年。[131] A. Hallur、T. Thebaud、G. Tinchev、V. Ravichandran 和 L. Moro-Velazquez，《面向口语对话系统的基于参考的韵律与节奏评估》，2026年。 [在线]。可访问：https: //arxiv.org/abs/ 2606 31055 。Y. A. Li、C. Han、V. Raghavan、G. Mischler 和 N. Mesgarani，《StyleTTS 2：通过风格扩散和大型语音语言模型对抗训练，实现接近人类水平的文本转语音》，《神经信息处理系统进展》，第36卷，第19621页，2023年，第19594页。–。[133] H. Viswanath、A. Bhattacharya、P. Jutras-Dub´e、P. Gupta、M. Prashanth、Y. Khaitan 和 A. Bera，《AffectEcho：面向语音合成的说话人无关、语言无关的情绪与情感迁移》，arXiv预印本 arXiv:2308.08577，2023年。[134] Y. A. Li、C. Han 和 N. Mesgarani，《SLMGAN：利用语音语言模型表示，在生成对抗网络中实现无监督零样本语音转换》，2023年。 [在线]。可访问：https: //arxiv.org/abs/ 2307 09435 。P. Tuttosı、S. Mehta、Z. Syvenky、B. Burkanova、G. E. Henter 和 A. Lim，《EmoJivoice：迈向机器人语音中长期可控的表现力》，2025年。 [在线]。可访问：https://arxiv.org/abs/2506.150 85[136] Y. Saito、E. Iimori、S. Takamichi、K. Tachibana 和 H. Saruwatari，《CALLS：客户中心投诉处理与专注倾听的日语共情对话语音语料库》，arXiv预印本 arXiv:2305.13713，2023年。[137] W. Zhang、X. Yang、Z. Gao、S. Feng、D. Wang 和 Y. Zhang，《PRISM：用于共情口语对话的韵律集成多智能体推理框架》，2026年。 [在线]。可访问：https: //arxiv.org/abs/ 2606 12902 。Z. Gao、X. Wang、X. Yang、W. Zhang、D. Wang、S. Feng 和 Y. Zhang，《ES4R：基于前置情感建模的语音编码，用于共情响应生成》，2026年。 [在线]。可访问：https://arxiv.org/abs/2601 16225 。Y. Hong、Y. Zhou 和 Y. Feng，《FreezeEmpath：利用冻结的大语言模型高效训练共情口语聊天机器人》，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2604 18159 。H. Liao、H. Han、X. Han、D. Chen、Y. Wang 和 Z. Wu，《CycleSpeech：用于指令控制语音合成和副语言理解的互惠对齐》，arXiv预印本 arXiv:2609.24771，2026年。 [ ] 2609 24771 在线。可访问：https://arxiv.org/abs/。S. Cheng、W. Bian、S. Cao、J. Li、B. Liu、C. Jiang、Y. Zhang、W. Wu、Y. Li、C.-M. Chan、L. Ma 和 W. Xue，《STEB：用于评估超越翻译保真度表现的语音到语音翻译表现力基准测试》，2026 [ ] 2606 25529 。在线。可访问：https://arxiv.org/abs/。A. Koudounas、H. Futami、Q. Jodelet、O. Take、S. Watanabe 和 E. Tsunoo，《语音到语音翻译模型的基准评测》，2026年。 [ ] 2606 03241 在线。可访问：https://arxiv.org/abs/。Y. Xue、S. Ouyang 和 L. Li，《长篇同步语音到语音翻译的实用评估方法》，2026年。 [在线]。可访问：https://arxiv.org/abs/2606 15059 。Y. Du、Q. Huang、G. Zhu、Z. Dai、S. Chen、Q. Zhu、L. Pan、M. Chen、Y. Zhang、L. Zhou、B. Wang 和 H. Li，《Mtalk-bench：通过竞技场式评测和评分量规协议评估多轮对话中的语音到语音模型》，2025年。 [在线]。可访问：https://arxiv.org/abs/2508.18240

[145] S.-C. Chen、I.-N. Tsai、Y.-C. Lin、S.-F. Huang 和 H. yi Lee，《Move：通过语音到语音翻译中的发声专家混合，实现笑声与哭声的翻译》，2026年。 [在线]。可访问：https://arxiv.org/abs/2604 17435 。Y. Pan、Y. Hou、X. Wu、L. Zhang、Y. L. Traon、L. Ma 和 J. Zhao，《从扁平语言标签到类型学先验：用于多语言语音到语音翻译的结构化语言条件》，2026 [ ] 2605 16026 。在线。可访问：https://arxiv.org/abs/。G.-T. Lin、J. Lian、T. Li、Q. Wang、G. Anumanchipalli、A. H. Liu 和 H. yi Lee，《全双工基准测试：评估全双工口语对话模型轮流发言能力的基准测试》，载于《2025年IEEE自动语音识别与理解研讨会（ASRU）》，2025， 1 8 [ ] 2503 04721 第页。在线。可访问：https://arxiv.org/abs/。Z. Zhao、S. Wang、G. Li、H. Xue、C. Wang、S. Wang、L. Xiao、Z. Zhang、H. Bu、X. Xu、X. Wang、H. Liu、E. S. Chng、H. yi Lee 和 L. Xie，《ICASSP 2026 HumDial 挑战赛：大语言模型时代类人语音对话系统的基准评测》，2026年。 [在线]。可访问：https://arxiv.org/abs/2601 05564 。C. Wang、H. Xue、G. Li、Z. Zhao、S. Wang、S. Wang、X. Xu、H. Bu 和 L. Xie，《口语对话系统中的全双工交互：基于 ICASSP 2026 HumDial 挑战赛的综合研究》，2026年。 [ ] 2604 21406 在线。可访问：https://arxiv.org/abs/。T. Thebaud、Y. Wang、H. Zhang、S. M. N. Ugandhar、A. Hallur、G. Tinchev、V. Ravichandran 和 L. Moro-Velazquez，《Spearbench：流式语音到语音语言模型自然度评估基准》，2026年。 [在线]。可访问：https: //arxiv.org/abs/2607 05365 。S. Arora、Z. Lu、C.-C. Chiu、R. Pang 和 S. Watanabe，《Talking Turns：音频基础模型轮流发言动态基准评测》，载于《国际学习表征会议》，第2025卷，2025年，第52754–52781页。[152] L. Lin、Y. Zhu、F. Zhang、Y. Liu、B. Li、X. Zeng、J. Wang 和 Y. Li，《MIRA：面向具身陪伴体的实时全双工人机交互》，arXiv预印本 arXiv:2609.24547，2026年。 [ ] 2609 24547 在线。可访问：https://arxiv.org/abs/。E. E. Shriberg，《语音不流畅现象理论初探》，博士论文，加州大学伯克利分校，1994年。[154] P. A. Heeman 和 J. Allen，《语音修正、语调短语与话语标记：口语对话中说话人话语的建模》，《计算语言学》，第25卷，第4期，第527–572页，1999年。[155] Y. Peng、Z. Ma、C. Liu、Y.-W. Chao、X. Chen 和 E. S. Chng，《面向不确定性的主动策略：口语对话系统的因果感知错误诊断与交互式澄清》，2026年。 [在线]。可访问：https://arxiv.org/abs/2605 25404 。K. Xu、Y. Wang 和 Y. Wang，《从被动响应到主动行动：通过 provoice-bench 评估语音智能体的主动性》，2026年。 [在线]。可访问：https://arxiv.org/abs/2604 15037 。G. Mialon、R. Dessı、M. Lomeli、C. Nalmpantis、R. Pasunuru、R. Raileanu、B. Rozi`ere、T. Schick、J. Dwivedi-Yu、A. Celikyil-maz 等，《增强型语言模型：综述》，arXiv预印本 arXiv:2302.07842，2023年。[158] Z. Xi、W. Chen、X. Guo、W. He、Y. Ding、B. Hong、M. Zhang、J. Wang、S. Jin、E. Zhou 等，《基于大语言模型的智能体的兴起与潜力：综述》，《中国科学：信息科学》，第68卷，第2期，第121101页，2025年。[159] S. Ultes，《利用双向长短期记忆网络改进交互质量估计及其对对话策略学习的影响》，载于《第20届SIGdial话语与对话年会论文集》，2019年，第11–20页。[160] D. Wu、T. Zhang、Y. Li、H. Liu、C. Chen、E. S. Chng 和 Y. Bengio，《静默思维：通过潜在推理对全双工口语对话模型中的内部认知进行建模》，2026年。 [在线]。可访问：https://arxiv.org/abs/2603 17837 。Y. Li、S. Ji、Y. Chen、T. Liang、H. Ying、Y. Wang、J. Li、J. Fang 和 Z. Zhao，《Wavbench：端到端口语对话模型的推理、口语化与副语言学基准评测》，2026年。 [在线]。可访问：https://arxiv.org/abs/2602 12135 。V. Karpukhin、B. Oguz、S. Min、P. Lewis、L. Wu、S. Edunov、D. Chen 和 W.-t. Yih，《面向开放域问答的密集段落检索》，载于《2020年自然语言处理实证方法会议论文集（EMNLP）》，2020年，第6769–6781页。[163] S. Borgeaud、A. Mensch、J. Hoffmann、T. Cai、E. Rutherford、K. Millican、G. B. Van Den Driessche、J.-B. Lespiau、B. Damoc、A. Clark 等，《通过从数万亿个词元中检索信息来改进语言模型》，载于《国际机器学习会议》。PMLR，2022年，第2206–2240页。[164] C.-M. Chien、M. Orsini、E. Kharitonov、N. Zeghidour、K. Livescu 和 A. D´efossez，《Moshirag：异步知识检索

2026年。全双工语音语言模型，https://arxiv.org/abs/2604 12928 。S. Arora、H. Khan、K. Sun、X. L. Dong、S. Choudhary、S. Moon、X. Zhang、A. Sagar、S. T. Appini、K. Patnaik、S. Sharma、S. Watanabe、A. Kumar、A. Aly、Y. Liu、F. Metze 和 Z. Lin，“Stream rag：借助流式工具使用实现即时且准确的口语对话系统”，2025 [ ] 2510 02044 。在线。可用地址：https://arxiv.org/abs/。T. Schick、J. Dwivedi-Yu、R. Dessı、R. Raileanu、M. Lomeli、E. Hambro、L. Zettlemoyer、N. Cancedda 和 T. Scialom，“Toolformer：语言模型可以自学使用工具”，神经信息处理系统进展，第36卷，第68 539–68 551页，2023年。[167] S. G. Patil、T. Zhang、X. Wang 和 J. E. Gonzalez，“Gorilla：连接海量 API 的大语言模型”，神经信息处理系统进展，第37卷，第126 544–126 565页，2024年。[168] W. Tan、X. Qu、M. Tu、M. Ge、A. T. Liu、P. Koehn 和 L. Lu，“面向交互式多模态工具使用智能体的过程监督强化学习”，2025 [ ] 2509 14480 。在线。可用地址：https://arxiv.org/abs/。A. Purwar 和 A. Choudhary，“Focal：一种面向多模态智能体的新型基准测试技术”，2026年。 [在线]。可用地址：https://arxiv.org/abs/2601 07367 。D. Mardien 和 F. Liu，“Mdwaistscheduler：一种用于免提临床排程的低成本语音激活设备”，2026年。 [在线]。可用地址：https://arxiv.org/abs/2604.16352[171] Yuma AI，“面向电子商务支持的语音 AI 智能体”，产品文档，2026年，访问于2026年9月。 [在线]。可用地址：https://yuma.ai/voice-ai-agent[172] ——，“Yuma AI 发布可在实时电话通话期间完成支持动作的语音智能体”，Yuma AI 产品公告，9月。 2026, 22 2026 [ ] 9月；早期测试版公告。在线。可用地址：https://yuma.ai/blogs/yuma-ai-launches-yuma-voice-ai [173] D. Driess、F. Xia、M. S. Sajjadi、C. Lynch、A. Chowdhery、B. Ichter、A. Wahid、J. Tompson、Q. Vuong、T. Yu 等，“Palm-e：具身多模态语言模型”，arXiv 预印本 arXiv:2303.03378，2023年。[174] A. Brohan、N. Brown、J. Carbajal、Y. Chebotar、J. Dabis、C. Finn、K. Gopalakrishnan、K. Hausman、A. Herzog、J. Hsu 等，“Rt-1：用于大规模现实世界控制的机器人变换器”，arXiv 预印本 arXiv:2212.06817，2022年。[175] A. Brohan、N. Brown、J. Carbajal、Y. Chebotar、X. Chen、K. Choro-manski、T. Ding、D. Driess、A. Dubey、C. Finn 等，“Rt-2：视觉-语言-动作模型将网络知识迁移至机器人控制”，arXiv 预印本 arXiv:2307.15818，2023年。[176] N. T. Walker、S. Ultes 和 P. Lison，“一种面向人机交互中知识 grounded 响应生成的图到文本方法”，2023 [ ] 2311 16137 。在线。可用地址：https://arxiv.org/abs/。J. Grannen、S. Karamcheti、S. Mirchandani、P. Liang 和 D. Sadigh，“Vocal sandbox：情境化人机协作中的持续学习与适应”，2024年。 [在线]。可用地址：https://arxiv.org/abs/2411 02599 。H. Kawai、D. Lala、K. Inoue、K. Ochi 和 T. Kawahara，“带有接管提示的半自主注意力倾听系统评估”，2024 [ ] 2402 14863 。在线。可用地址：https://arxiv.org/abs/。M. Savva、A. Kadian、O. Maksymets、Y. Zhao、E. Wijmans、B. Jain、J. Straub、J. Liu、V. Koltun、J. Malik 等，“Habitat：用于具身人工智能研究的平台”，载于2019年 IEEE/CVF 国际计算机视觉会议（ICCV）。IEEE，2019年，第9338–9346页。[180] M. Shridhar、J. Thomason、D. Gordon、Y. Bisk、W. Han、R. Mottaghi、L. Zettlemoyer 和 D. Fox，“ALFRED：用于解释日常任务中基于环境的指令的基准测试”，载于2020年 IEEE/CVF 计算机视觉与模式识别会议（CVPR）。IEEE，2020年，第10 737–10 746页。[181] K. Wang、H. Ren、Z. Lu、M. Zhan 和 H. Li，“VoiceAssistant-Eval：对人工智能助手在听觉理解、语音表达和视觉观察方面进行基准评测”，2025 [ ] 2509 22651 。在线。可用地址：https://arxiv.org/abs/。M. Chen、X. Zhang、M. Peng、Z. Yu、A. Papangelis 和 Y. Jo，“MIST：面向智能家居的多模态交互式基于语音的工具调用对话助手”，2026年。 [在线]。可用地址：https://arxiv.org/abs/2605 06897 。T. Chen、B. Veluri、H. Gong 和 S. Gollakota，“AV-Dialog：带有视听输入的口语对话模型”，2025年。 [在线]。可用地址：https://arxiv.org/abs/2511 11124 。J. Jiang、A. Ohashi 和 R. Higashinaka，“将面部生成整合至全双工口语对话系统中”，2026年。 [在线]。可用地址：https://arxiv.org/abs/2606 21970 。T. Liang、Y. Chen、S. Ji、Y. Chen、Z. Jia、J. Lu、F. Zhuo、X. Pu、Y. Li 和 Z. Zhao，“VoxMind：一种端到端智能体式口语对话系统

，” 2026年。 [在线]。可访问：https://arxiv.org/abs/2604.15710

[186] Z. Wan、C.-H. H. Yang、J. Tian、H. Ye、A. Pasad、S. wei Fu、A. Goel、R. Hachiuma、S. Diao、K. Dhawan、S. Ghosh、Y. Hirota、Z. Chen、R. Valle、C. Chu、S. Watanabe、Y.-C. F. Wang 和 B. Ginsburg，“Speech-hands：一种用于语音识别和全方位感知音频推理的自我反思式语音智能体方法”，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2601 09413 。J. Kirmayr、L. Stappen 和 E. Andre，“Car-bench：评估大语言模型智能体在现实世界不确定性下的一致性与边界意识”， 2026 [ ] 2601 22027 。[187] 在线。可访问：https://arxiv.org/abs/。J. Kirmayr、R. Wennmacher、K. Huynh、L. Stappen、E. Andre 和 F. Alt，“你在做什么？多步处理期间来自智能体大语言模型车载助手的中间反馈的影响”，2026年。 [ ] 2602 15569 在线。可访问：https://arxiv.org/abs/。T. Bogavelli、G. G. Melanc,on、K. Stankiewicz、O. Bamgbose、F. Riols、H. H. Nguyen、R. Mehndiratta、L. D. Brin、J. Marinier、H. Subramani、A. Madamala、S. K. Nemala 和 S. Sunkara，“Eva-bench：用于评估语音智能体的新型端到端框架”，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2605 13841 。M. Todisco、X. Wang、V. Vestman、M. Sahidullah、H. Delgado、A. Nautsch、J. Yamagishi、N. Evans、T. Kinnunen 和 K. A. Lee，“ASVspoof 2019：伪造语音和合成音频检测的未来方向”，arXiv预印本 arXiv:1904.05441，2019年。[191] N. Carlini、F. Tramer、E. Wallace、M. Jagielski、A. Herbert-Voss、K. Lee、A. Roberts、T. Brown、D. Song、U. Erlingsson 等，“从大语言模型中提取训练数据”，载于第30届USENIX安全研讨会（USENIX Security 21），2021年，第2633–2650页。[192] N. Carlini 和 D. Wagner，“音频对抗样本：针对语音转文本的定向攻击”，载于2018年电气与电子工程师协会安全与隐私研讨会（SPW）。电气与电子工程师协会，2018年，第1–7页。[193] L. Sch¨onherr、K. Kohls、S. Zeiler、T. Holz 和 D. Kolossa，“通过心理声学隐藏攻击自动语音识别系统”，arXiv预印本 arXiv:1808.05665，2018年。[194] M. Kang、C. Fang 和 B. Li，“Audioguard：面向多样化威胁模型的全面音频安全防护”，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2604 08867 。Y. Lu、Y. Yan、Z. Zhang、Z. Ou、R. Liu、L. Liu 和 S. Wang，“语音令牌会泄露声纹吗？针对端到端语音语言模型的说话人反演攻击”，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2607 16870 。S. Cheng、K. Xiang、M. Li、J. Zhang、D. Liu 和 W. Jiang，“永不停下语音：针对端到端语音语言模型的拒绝服务攻击”，2026 [ ] 2608 10405 。在线。可访问：https://arxiv.org/abs/。K. Greshake、S. Abdelnabi、S. Mishra、C. Endres、T. Holz 和 M. Fritz，“并非你报名时所期待的：通过间接提示注入破坏现实世界中的大语言模型集成应用程序”，载于第16届ACM人工智能与安全研讨会论文集，2023年，第79–90页。[198] R. Chen、W. Liang、Z. Gong、L. Ai 和 J. Hirschberg，“通过合成多说话人对话检测语音中的心理操纵”， 2026 [ ] 2601 08342 。在线。可访问：https://arxiv.org/abs/。K. Miller、A. Chandra 和 V. Saligrama，“音频语言模型会使用副语言证据吗？用于响应评估的反事实审计”，2026 [ ] 2608 06718 。在线。可访问：https://arxiv.org/abs/。A. Purwar、S. Singh 和 K. Srivastava，“大语言模型评审器基准评测：语音智能体评估的可靠性、校准与人工监督”，2026 [ ] 2608 24314 。在线。可访问：https://arxiv.org/abs/。T. Kawahara、K. Inoue 和 D. Lala，“应用于专注倾听和求职面试的智能对话机器人Erica”，arXiv预印本 arXiv:2105.00403，2021年。[202] L. Brinkschulte、N. Mariacher、S. Schl¨ogl、M. I. Torres、R. Justo、J. M. Olaso、A. Esposito、G. Cordasco、G. Chollet、C. Glackin 等，“The empathic项目：构建富有表现力的先进虚拟教练，以改善老年人的独立健康寿命年”，arXiv预印本 arXiv:2104.13836，2021年。[203] M. I. Torres、J. M. Olaso、C. Montenegro、R. Santana、A. V´azquez、R. Justo、J. A. Lozano、S. Schl¨ogl、G. Chollet、N. Dugan 等，“The empathic项目：中期成果”，载于第12届ACM辅助环境相关普适技术国际会议论文集，2019年，第629–638页。 [在线]。可访问：https://arxiv.org/abs/ 2105 01878 。Y. Bai、J. Chen、Y. Chen、W. Deng、B. Dong、N. Duan、Y. Gu、W. Han、Y. Huang、M. Ke、H. Li、J. Li、X. Liang、N. Liu、Y. Liu、J. Miao、J. Wang、Q. Wang、W. Wang、Y. Wang、Z. Wang、Z. Xiao、C. Xue、H. Xue、F. Yu、T. Zhang、Y. Zhang、Y. Zhang 和 L. Zhu，“Joyai-talker：为共情语音智能体打造的全双工语音交互大模型”，2026 [ ] 2608 01119 。在线。可访问：https://arxiv.org/abs/。J. D. Lee 和 K. A. See，“自动化信任：为恰当依赖进行设计”，Human factors，第46卷，第1期，第50–80页，2004年。[204] [205] [206] K. A. Hoff 和 M. Bashir，《自动化信任：整合影响信任因素的实证证据》，*Human factors*，第 57 卷，第 3 期，第 407–434 页，2015年。[207] A. Mahmood、B. Kim、H. Zhao、M. E. Atwood、L. F. Buenaver、M. T. Smith 和 C.-M. Huang，《更高的依从性，更丰富的语境：面向睡眠的大语言模型驱动的对话式语音日记的实地评估》，2026年。 [ ] 2606 18596 在线。可访问：https://arxiv.org/abs/ 。N. Mathur、H. Rahman 和 S. Desai，《宜人性和外向性对辅助场景中老年人对对话式人工智能解释的感知所产生的差异化影响》，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2603 08164 。P. Ravi、C. Stevens、B. Hurt、B. Hanks、G. Lin 和 E. Anderson，《探索智能体语音口音如何塑造 k-12 2026 [ ] 26 群体学习中的人机协作》。在线。可访问：https://arxiv.org/abs/ 。[210] A. Koenecke、A. Nam、E. Lake、J. Nudell、M. Quartey、Z. Mengesha、C. Toups、J. R. Rickford、D. Jurafsky 和 S. Goel，《自动语音识别中的种族差异》，《美国国家科学院院刊》，第 117 卷，第 14 期，第 7684–7689 页，2020年。[211] A. C. Kocabiyikoglu、F. Portet、J.-M. Babouchkine、P. Gibert、H. Blanchon 和 G. Gavazzi，《用于医疗处方获取的口语对话系统：开发、语料库与评估》，2023 [ ] 2311 03510 。在线。可访问：https://arxiv.org/abs/ 。Y. Ni 和 F. Jia，《心理健康护理中人工智能驱动的数字干预范围综述：绘制筛查、支持、监控、预防和临床教育中的应用》，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2603 16204 。A. Addlesee 和 P. Albert，《与认知障碍人士开展多模态自发对话的伦理收集》，*arXiv* 预印本 *arXiv:2009.14361*，2020年。[214] H. Suresh 和 J. Guttag，《理解机器学习生命周期中危害来源的框架》，载于*ACM*算法、机制与优化公平性和可及性会议第一届会议录，2021年，第 1–9 页。[215] E. Debenedetti、J. Zhang、M. Balunovic、L. Beurer-Kellner、M. Fischer 和 F. Tram`er，《Agentdojo：用于评估大语言模型智能体提示注入攻击与防御的动态环境》，神经信息处理系统进展，第 37 卷，第 82 895–82 920 页，2024年。[216] I. D. Raji、A. Smart、R. N. White、M. Mitchell、T. Gebru、B. Hutchinson、J. Smith-Loud、D. Theron 和 P. Barnes，《弥合人工智能问责鸿沟：定义端到端内部算法审计框架》，载于*2020*年公平性、问责制与透明度会议录，2020年，第 33–44 页。[217] Y. Chiba 和 R. Higashinaka，《研究增量处理和语音活动预测对口语对话系统的影响》，载于第 *31* 届国际计算语言学会议录，2025年，第 3687–3696 页。[218] K. Qi、Y. Zhu、C. Summerour、J. A. Batsis 和 X. Liang，《Cog-tipro：利用大语言模型进行迭代式提示优化，通过纵向语音助手指令检测认知衰退》，载于*GLOBECOM 2025-2025*年*IEEE*全球通信大会。电气与电子工程师协会，2025年，第 356–362 页。[219] E. Fruits 和 K. Stout，《人工智能、生产率与劳动力市场：实证证据综述》，*SSRN 6323960* 可获取，2026年。