湖南大学

HUNAN UNIVERSITY

本科生毕业论文(设计)

论文(设计)题目：	低延迟响应与自然交互的 全双工对话系统研究
学生姓名：	于浩源
学生学号：	202208010230
专业班级：	计算机科学与技术 2202
学院名称：	计算机学院
指导老师：	蔡敏捷

2026 年 04 月 30 日

湖南大学

毕业论文（设计）原创性声明

本人郑重声明：所呈交的论文（设计）是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写的成果作品。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律后果由本人承担。

学生签名：　　日期：　　年　月　日

毕业论文（设计）版权使用授权书

本毕业论文（设计）作者完全了解学校有关保留、使用论文（设计）的规定，同意学校保留并向国家有关部门或机构送交论文（设计）的复印件和电子版，允许论文（设计）被查阅和借阅。本人授权湖南大学可以将本论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本论文（设计）。

本论文（设计）属于

1、保　密□，在_____年解密后适用本授权书。

2、不保密□。

（请在以上相应方框内打“√”）

学生签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

低延迟响应与自然交互的全双工对话系统研究

摘　　要

随着大语言模型，多模态模型与语音语言模型的快速发展，语音对话系统正由传统轮次式交互逐步迈向更加接近人类交流方式的同步式交互。现有多数语音系统仍采用监听，理解，生成的顺序流程，在处理自然对话中的重叠发言，附和，打断和非线性交互时存在明显局限。难以同时满足低延迟响应与自然交互这两个关键目标。尤其在真实场景下，用户可能随时打断系统，重复表述，短暂停顿后继续发言，甚至存在背景第三方语音干扰，这些情况都会对系统的状态切换，说话权管理和响应合理性提出更高要求。因此，围绕低延迟响应与自然交互开展全双工对话系统研究，具有重要的理论意义和应用价值。

本文以全双工对话系统为主线，围绕系统设计，状态管理与音频理解能力展开研究整理。首先对全双工系统进行了形式化区分，总结了当前全双工语音语言模型在体系结构与评测方法上的研究进展，为本文系统设计提供理论基础。其次，针对传统级联系统副语言信息利用不足，端到端全双工模型训练成本过高的问题，我们提出了一种基于对话单元的半级联全双工语音对话框架。该框架以多模态大语言模型为核心，通过两类控制动作管理监听与说话两种状态，在保持可插拔工程结构的同时提升了对打断，附和和低时延交互的支持能力。最后，围绕系统中的音频理解模块，我们将 UniWhisper 所代表的统一音频表示方法作为一种扩展性探索方案加以分析，讨论其在增强全双工系统音频感知能力，提升复杂场景鲁棒性方面的潜在价值。

相关领域数据集上的实验结果表明，我们提出的系统模块与整体框架在低延迟响应和自然交互方面具有明显有效性，验证了其在时序建模，状态管理与音频理解等关键环节的协同作用。

关键词： 全双工对话系统；人机交互；多模态大语言模型；音频理解

Full-Duplex Dialogue Systems for Low-Latency Response and Natural Interaction

Abstract

Recent advances in large language models, multimodal models, and speech language models are driving spoken dialogue systems from turn-based interaction toward more human-like synchronous interaction. However, most existing systems still rely on a sequential listen-understand-generate pipeline, which struggles with overlapping speech, backchannels, interruptions, and other non-linear conversational behaviors. This makes it difficult to achieve both low latency and natural interaction, especially in real-world settings.

This thesis studies full-duplex dialogue systems from the perspectives of system design, state management, and audio understanding. We first formalize the distinction of full-duplex systems and review recent progress in full-duplex speech language models. We then propose a dialogue-unit-based semi-cascaded full-duplex spoken dialogue framework, which improves support for interruptions, backchannels, and low-latency interaction while maintaining a modular design. Finally, we analyze unified audio representation methods, exemplified by Uni-Whisper, as a potential extension for improving audio perception and robustness in complex scenarios.

Experiments on relevant datasets demonstrate the effectiveness of the proposed framework in achieving low-latency response and natural interaction.

Keywords: **full-duplex dialogue system; human computer interaction; multimodal large language model; audio understanding**

目　　录

插图索引

附 表 索 引

1 绪论

1.1 研究背景

近年来, 以大语言模型[1]为代表的人工智能技术快速发展, 推动了自然语言理解[2], 知识推理[3], 内容生成和多轮对话系统的持续演进[4]。在这一背景下，语音交互系统也由传统的语音识别[5], 自然语言处理[6], 语音合成[7]的级联式方案，逐步迈向更统一的语音语言模型框架[8]。尤其是在多模态大模型[1]出现之后，系统不再仅将语音视作文本的载体，而是尝试直接从连续音频流中提取语义，韵律和说话人状态等丰富信息。然而，当前大量语音系统仍然遵循用户先说完，系统再完整回复的轮次式模式，即使其语音理解能力和音频生成质量已经明显提升，但交互形式与交互体验仍与人类自然对话存在显著距离。

在人与人之间的日常交流中，说话和倾听并不是截然分开的两个线性阶段，而是彼此交叠, 相互预测并动态协商的过程。对话双方会根据语气，停顿，语义完整度和话轮趋势判断是否应该继续发言，附和或打断。例如，在面对熟悉话题时，听者往往会用简短的附和语表达正在跟随。而在意识到对方误解自己的意图时，又可能立即插话进行修正。传统半双工系统由于同一时刻只允许单向音频流存在，无法支持这一类自然行为。即使某些系统通过快速切片，缓冲和调度制造出边听边说的效果，其核心仍然是顺序处理，而不是真正意义上的同步交互。因此，构建能够并行完成输入感知与输出生成的全双工语音语言模型，成为语音智能体迈向类人交互的关键一步。

与传统文本对话不同，语音交互不仅涉及说话内容，还涉及说话方式，恰当的说话时间。自然对话中的重叠发言, 迟疑停顿, 情绪波动和说话节奏等副语言线索，往往对系统理解用户意图, 判断说话权转移时机具有决定性作用。与此同时，真实应用环境还常常伴随复杂的声学事件，如背景噪声，第三方说话人乃至音乐片段。也就是说，一个真正可用的语音智能体，不能只具备将语音转写为文本的能力，还需要形成对多种音频模态的统一而稳健的表征。如果底层编码器过度偏向语音转写任务，则可能忽视环境声和全局语义。如果过度偏向一般音频语义，又可能损失语音任务所需的细粒度时序信息。当前语音智能体研究一方面在交互层面探索全双工建模，希望系统具备边听边说、及时打断和灵活轮换的能力，另一方面也在关注音频理解能力本身是否足以支撑复杂状态判断。前者直接对应低延迟响应与自然交互的系统目标，后者更多作为支撑系统表现的底

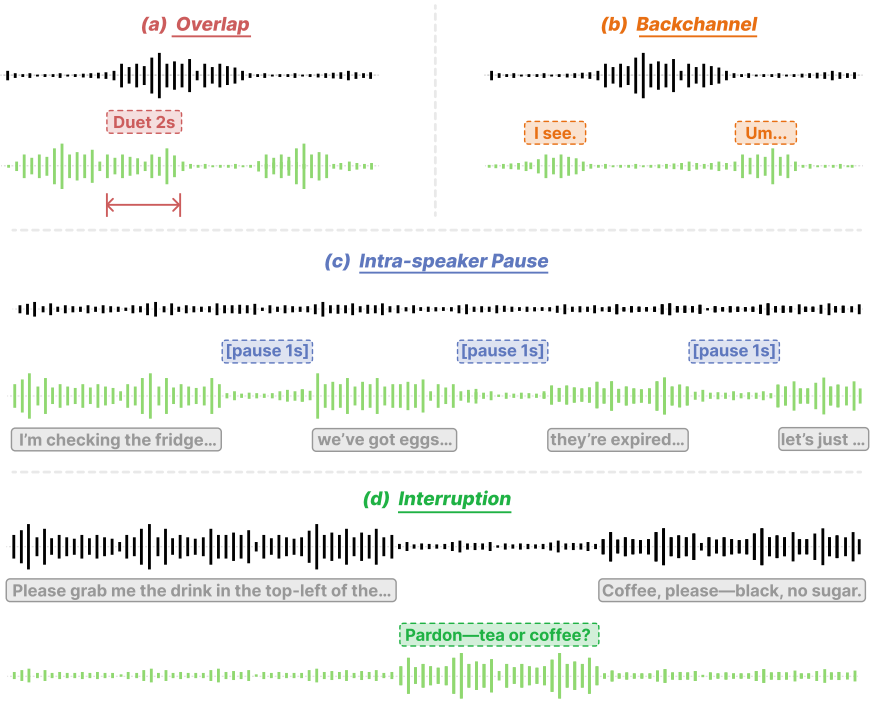

图 1.1 自然对话中存在的情况：（a）重叠，（b）附和，（c）暂停，（d）打断。

层能力。两者共同决定了未来全双工对话系统能否真正适应复杂真实场景中的长时连续的开放式交互。

1.2 研究意义

从理论意义上看，对全双工语音语言模型的整理与分析，有助于明确语音交互系统从轮次式处理走向同步式处理的技术演化逻辑。传统研究往往将全双工能力视作系统实现层面的附加功能，而较少将其上升为一个需要形式化界定和统一评测的核心建模问题。通过对半双工，伪全双工和真正全双工的严格区分，可以更清晰地理解不同系统之间在认知并行性，时序建模和控制机制上的本质差别。通过对工程化同步与学习式同步两类技术路线的梳理，可以把当前全双工研究从零散个案总结为较为明确的研究系统。通过进一步引入统一音频表示这一基础能力视角，则能够将语音智能体研究从如何对话延伸到如何感知，从而形成自上而下更加完整的理论框架。

从方法论意义上看，我们搭建了系统交互机制和音频理解能力之间的联系。全双工对话系统之所以难以实现稳定自然的听说切换，一个重要原因就在于系统往往无法准确把握输入音频中的副语言线索和场景信息。而统一音频表示学习可以作为音频理解模块的一种潜在增强路线，为这一问题提供底层支持。将两者放在同一研究脉络中加以讨

论，能够帮助我们从系统工程和基础模型两个维度理解自然语音交互的核心瓶颈。

从应用价值上看，具备低时延，可打断，可附和，可持续交互等功能的语音系统，在智能助手，车载语音，人机协作，陪伴机器人，智能教育与无障碍交互等场景中均具有广阔前景。对于这类系统而言，用户并不满足于能听懂，能回答，而更强调系统能否会在恰当时机接话，在被打断后快速恢复以及在噪声环境中保持稳定理解。特别是在开放场景中，如果系统既能保持稳定的话轮转换，又能对环境声，情绪变化和说话状态形成较全面感知，其交互自然性，鲁棒性和用户满意度都将显著提高。因此，围绕全双工对话系统展开研究，不仅具有明确的学术价值，也具有现实的工程应用意义。

1.3 研究现状

1.3.1 全双工语音语言模型研究现状

近年来出现了一批面向全双工交互的语音语言模型工作。当前方法主要遵循两种范式，一是通过模块化架构实现的工程化同步，二是通过端到端系统学习得到的同步机制。工程化同步机制主要通过引入专门的功能模块来增强对话系统的全双工能力，其核心思想是在不重新训练主模型的前提下，通过显式的状态机制实现对交互行为的控制。

工程化同步机制通常可以分为外部控制器和内部控制器两类。外部控制器通过独立于主对话引擎的控制模块来决定系统是否应当继续发言或切换状态，其特点是结构清晰且即插即用，便于与既有系统解耦。例如，FlexDuo[9] 设计了一个包含空闲状态的三态有限状态机，用于实现选择性注意力控制。Semantic VAD[10] 则利用轻量级模型分析自动语音识别结果，以降低计算开销。Phoenix-VAD[11] 提出一种可插拔的 LLM 驱动语义端点检测模块，通过滑动窗口方式直接对语音流进行流式语义完结性判断，在无需额外 ASR 的情况下区分继续说话与停止说话，从而为全双工系统提供独立可部署的外部控制能力。SoulX-Duplug[12] 提出一种可插拔的流式状态预测模块，将 VAD、流式 ASR 与轮次状态预测统一到单一框架中，并通过文本引导的状态 token 预测与教师强制式 ASR 推理实现低时延、语义感知的全双工对话控制。VITA-1.5[13] 采用双实例交替工作的方式，在检测到打断时进行角色切换，从而以一定的计算代价换取更低的交互延迟。FireRedChat[14] 构建了一套完整的可插拔全双工语音交互系统，通过流式个性化语音活动检测与语义级轮次结束检测组成的外部控制器，在不修改主模型的前提下将任意半双工流水线升级为全双工交互，并在虚假打断率与端到端延迟上均接近工业级系

统水平。内部控制器则将控制逻辑直接嵌入到对话引擎内部，使状态管理成为模型结构的一部分。例如，Freeze-Omni[15] 在冻结的大语言模型上进行分块状态预测，实现对交互时序的控制；MinMo[16] 的 Full Duplex Predictor 根据输入嵌入判断何时让出发言权；Neural-FSM[17] 通过扩展有限状态机标记，使模型能够借助下一词预测自动完成状态管理；Mini-Omni2[18] 则利用语义状态标记实现基于指令的打断控制。DuplexMamba[19] 通过引入基于状态 token 的双工解码策略，在 Mamba 架构中为输入流显式判别需响应，未完成或可忽略等状态，并借助分支复制与切换机制实现流式输入处理、打断响应与非唤醒过滤。RTTL-DG[20] 将对话管理与响应生成统一到实时无文本框架中，通过每 160ms 预测保持静默、开始说话、继续说话或停止说话等动作，并直接生成语音单元，从而在不依赖文本中间表示的情况下实现更低延迟、更自然的重叠、打断与附和反馈。总体来看，这类方法的优势在于控制行为明确、可解释性较强，但其性能往往依赖于额外模块的设计质量，以及状态切换策略是否足够稳健。

与工程化同步机制不同，端到端同步机制直接对双向音频流进行建模，不再依赖显式的外部控制模块。自 dGSLM[21] 证明原始音频中可以自然涌现轮次切换行为以来，这类方法通常将全双工能力内化为模型本身的学习目标。其主要挑战在于如何协调 Transformer 的序列生成特性与对话交互中的并行处理需求，因此这类方法通常会在模态接口和流式处理方式上进行进一步设计。

从模态接口来看，不同方法在输入输出表示上存在明显差异。基于编解码器的方法通常将音频离散化为一系列 token，尽管这种方式会带来序列长度增加的问题，但便于与语言模型结合，如 dGSLM[21]、Moshi[22]、SyncLLM[23]、NTPP[24] 和 Voila[25]。SALMONN-omni[26] 则直接处理连续音频嵌入，在建模方式上更加接近原始声学空间；SALM-Duplex[27] 则结合连续输入与离散输出，在识别精度与响应时延之间进行折中。Covo-Audio[28] 同样采用连续输入与离散输出的混合双流方案，并将全双工能力直接内化于大规模预训练阶段，从而无需细粒度文本-语音对齐即可获得鲁棒的全双工对话能力。从流式处理方式来看，这类方法大体可以分为多流和单流两种范式。多流方法通常采用语音文本双流结构，并通过交叉注意力机制实现不同流之间的同步，如 dGSLM[21]。Moshi[22] 则基于 RQ-Transformer 对用户音频、系统输出以及内部独白进行联合建模，从而显式建模多路信息的协同关系。Fun-Audio-Chat-Duplex[29] 同样采用离散语音 token，通过引入并行语音文本输入流架构支持全双工对话，并以双分辨率语音表示，在主干

LLM 的 5Hz 高效处理与语音编码器的 25Hz 高质量生成之间取得平衡。单流方法则将输入序列化后交由标准解码器处理，例如 SyncLLM[23] 通过插入同步标记实现音频块间的交替处理。NTPP[24] 采用成对因果掩码进行联合预测，而 LSLM[30] 和 SALM-Duplex[27] 则探索了不同层次的特征融合方式。OmniFlatten[31] 通过多阶段渐进式后训练将文本 LLM 逐步适配为全双工对话模型，以固定块大小对语音与文本流进行分块后展平为单一序列，实现实时流式全双工交互。

在交互建模上，现有方法大多采用隐式交互方法，即通过模型生成静音、语音或特定控制 token 来实现轮次切换，而无需对这些行为进行显式标注或监督，如 dGSLM[21], Moshi[22], SyncLLM[23], NTPP[24] 和 Voila[25]。与此相比，SALMONN-omni[26] 提出的 Dynamic Thinking 机制通过生成控制 token 实现显式状态管理，从而将大语言模型定位为端到端系统中的全双工预测器。UAF[32] 则进一步在同一大语言模型中并行设计 LM Head, VAD Head 与 Turn Head 三个解码器头，通过输出控制 token 对语音活动检测与轮次状态进行显式联合建模，从而将全双工前端感知直接内化于语言模型的生成过程中。总体而言，端到端同步机制更强调系统的一体化建模能力，其优势在于能够从数据中自动学习交互规律，但也对模型结构, 训练数据以及推理效率提出了更高要求。

1.3.2 统一音频表示研究现状

统一音频表示的发展在很大程度上体现为音频预训练范式从底层声学建模向更高层语义建模不断演进的过程。wav2vec 2.0[33] 奠定了基于自监督学习的语音表示预训练基础，其核心思想是通过掩码预测和对比学习在大规模无标注语音中学习具有迁移能力的时序表示，使模型能够从原始波形中提取稳定的局部声学结构和上下文信息，这一工作推动了音频表示研究从依赖人工标注的监督学习转向依赖海量无标注数据的预训练学习。HuBERT[34] 在此基础上进一步发展出基于聚类伪标签的掩码预测框架，通过预测隐藏单元标签而非直接进行对比判别，使模型能够更有效地捕获语音中的潜在结构和语言相关信息，也使音频表征学习从关注连续声学信号本身进一步转向关注其中更抽象的离散结构。WavLM[35] 则延续了 HuBERT[34] 的基本思路，并通过在预训练中引入更复杂的语音场景和多种语音属性相关信息，使模型学习到同时包含语音内容特征，说话人特征和场景鲁棒性的统一语音表示，这表明语音基础模型的发展已经不再局限于单一的语音识别目标，而是开始朝向能够支撑多种语音任务的通用表征方向演进。与上述以编码表示学习为核心的模型不同，Whisper[36] 进一步体现出另一条重要的发展路径，即

通过大规模弱监督语音文本数据训练端到端的编码器解码器模型，将语音识别，语音翻译和语言识别等任务统一到同一生成框架下，使音频建模从单纯追求可迁移表示扩展到统一任务接口和统一输出空间的层面。总体来看，从 wav2vec 2.0[33] 到 HuBERT[34] 再到 WavLM[35] 和 Whisper[36]，统一音频表示研究逐步形成了从自监督声学表示学习，到基于离散单元的结构建模，再到面向多任务统一表征和生成式建模的发展脉络，反映出该领域的核心目标正在从学习通用语音特征进一步转向构建能够适应更广泛任务形式和更复杂应用场景的统一音频基础模型。当前这一方向的研究仍面临不同任务目标和不同音频类型之间统一建模能力不足的问题，如何在统一框架下进一步提升表示的泛化性和可迁移性，仍然是该领域的重要研究方向。

1.4 设计及研究内容

围绕低延迟响应与自然交互这两个全双工语音交互中的核心问题, 我们从系统构建, 音频理解增强和实验验证三个方面展开研究。首先, 针对真实交互过程中用户打断, 重叠发言, 重复输入以及背景第三方语音等现象容易造成系统等待增加, 状态切换混乱和交互不连贯的问题, 构建了一个由语音活动检测, 自动语音识别, 多模态大语言模型和语音合成模块组成的全双工语音对话系统[37], 并设计了相应的状态管理机制, 以协调各模块在连续交互过程中的行为, 提高系统在复杂场景下的稳定性与可用性。其次, 针对系统在自然交互中对复杂音频信号利用不足, 音频理解能力受限的问题, 提出了 UniWhisper[38] 持续多任务训练框架, 通过统一指令格式组织多种音频任务, 提升音频编码器的统一表示能力, 并围绕该框架开展了大规模实验, 系统验证了其在多任务音频理解场景中的有效性。最后, 针对所构建的全双工语音对话系统开展了大规模实验评测, 在 Full-Duplex-Bench v1.5[39] 和 ICASSP 2026 HumDial Challenge[40] 等评测设置下, 对停止延迟, 响应延迟, 恢复与拒识表现以及行为合理性等指标进行了系统分析。实验结果表明, 我们所构建的系统在真实全双工语音交互场景中具有良好的有效性, 并在 ICASSP 2026 HumDial Challenge[40] 的 Full-Duplex Interaction 任务中取得第二名的成绩。

具体而言, 在低延迟响应方面, 我们重点从系统架构, 推理优化和模块轻量化三个层面进行了设计。针对传统级联式语音对话系统中模块严格串行执行所带来的额外时延, 我们对全双工对话过程进行了建模, 设计了更适合连续交互的轮次转换机制, 使系统能够从传统级联方式转向半级联方式, 从而支持关键模块并行运行并减少跨模块传递造成的延迟损耗, 这是我们在系统实时性方面的核心创新。与此同时, 为了进一步降低部署

过程中的实际推理时延, 我们围绕关键模块开展了基础设施适配, 结合 vLLM 等推理加速手段对模型服务和异步执行流程进行了系统优化。此外, 针对多模态大语言模型计算开销较大的问题, 我们还探索了可替代其部分功能的轻量化模块, 以在保证交互能力的同时进一步提升系统响应效率。在自然交互方面, 我们分别从听和说两个角度进行探索。在输入理解侧, 系统主要依赖多模态大语言模型完成对复杂语音输入和场景信息的通用理解, 同时我们也进一步探索了更轻量化的替代方案, 以降低对大规模多模态模型的依赖。作为这一方向上的研究, 我们提出了 UniWhisper[38] 持续多任务训练框架, 通过统一音频表示学习提升模型对语音内容和更广泛音频信息的建模能力, 并用于验证轻量化音频理解方案在自然交互场景中的潜在价值。在输出生成侧, 我们引入了更具情感表现力和自然表达能力的语音合成模块, 使系统生成的回复在韵律和表达方式上更接近自然人机对话。整体而言, 我们的研究工作围绕低延迟响应与自然交互两条主线展开, 试图从系统效率和交互体验两个方面共同推动全双工语音对话系统向更自然, 更实用的方向发展。

我们的主要工作可以概括为以下三点：

1. 设计并实现了一个面向自然语音交互的全双工对话系统[37], 通过状态管理机制提升了系统对打断, 重复发言和背景第三方语音等复杂情况的处理能力。

2. 提出了 UniWhisper[38] 持续多任务训练框架, 并围绕该框架开展了大规模实验, 验证了其在多任务音频理解中的有效性。

3. 进行了面向真实全双工语音交互场景的大规模实验评测, 验证了所提出系统设计的有效性。

2 相关理论与技术

2.1 全双工语音语言模型的基本概念

全双工语音语言模型的目标是在同一交互过程中联合建模输入语音流与输出语音流的同步演化，使系统能够在持续接收用户语音的同时生成自身语音，并根据新增观测实时调整生成行为。其关键在于感知状态更新与响应生成在时间轴上的并行推进。全双工系统允许输入与输出在时间上发生重叠。全双工模型在持续接收输入的同时维持输出生成，并对当前输出进行在线修正延续或中止。

从形式化角度看，全双工语音语言模型需要刻画输入语音序列输出语音序列以及交互状态随时间的联合演化关系。由于原始语音是连续时间信号，而实际建模通常建立在离散单元声学片段或分帧连续表示之上，因此通常先将双向语音流映射为可对齐的序列表征，再定义条件生成或联合生成过程。这类模型的本质特征在于显式建模说话权转移重叠发声与在线反馈所对应的时序耦合关系。

2.2 全双工语音语言模型的架构分类

当前全双工语音语言模型主要可分为工程化同步与学习式同步两大范式。

2.2.1 工程化同步

工程化同步是在既有语音对话引擎之上引入显式同步机制的技术路线，通常建立在自动语音识别模块，大语言模型模块和文本转语音模块组成的模块化流水线之上，并通过额外控制单元协调输入处理与输出生成。该路线将全双工能力建模为显式的状态控制问题，系统架构通常由对话核心模型与同步控制模块共同构成。按照控制逻辑所在位置的不同，工程化同步可进一步分为外部控制器和内部控制器两种形式。外部控制器独立于主干模型，常见实现包括有限状态机、语义语音活动检测模型、端点检测器和双实例切换策略。内部控制器则将同步逻辑嵌入模型内部，典型形式包括状态预测头、特殊控制 token、神经有限状态机和分块级状态判别单元。总体上，这一路线的架构特征在于以模块解耦的方式组织全双工能力，并通过显式状态空间和控制策略实现输入流和输出流之间的协调。

2.2.2 学习式同步

学习式同步是通过统一参数化模型直接建模双向语音流的技术路线，其核心特征是将输入语音流与输出语音流纳入同一生成框架之中。该路线在表示层面可采用离散编解码器 token、连续声学表征或混合表示，在结构层面可采用双流架构、单流串行化架构或联合自回归架构。离散表示方法通常先借助神经编解码器将连续语音映射为 token 序列，再利用语言模型式结构进行建模。连续表示方法则直接处理分帧声学嵌入，并通过流式编码器、交叉注意力、掩码机制或专门的同步单元组织时序依赖。对应地，双流架构分别维护用户流和系统流，并通过跨流交互完成同步。单流架构将双向信号串行化到统一序列中，并借助同步 token 或因果掩码维持对齐关系。联合自回归架构则在同一生成过程中直接描述输入输出序列的联合演化。总体上，这一路线的架构特征在于以统一表示和统一生成过程组织全双工能力。

2.3 全双工语音语言模型的形式化描述

全双工语音语言模型通过同时对输入进行编码并对输出进行解码，实现对语音交互过程的并行建模，从而支持低延迟可实时调整的交互输出。设智能体为 $\mathcal{A}$，交互环境为 $\mathcal{E}$。由于连续音频信号 $X(t)$ 难以直接建模，通常首先通过离散化映射 $\mathcal{T}$ 将其转换为对齐的符号序列，记为环境序列 $S^{\mathcal{E}} = (e_1, \ldots, e_T)$ 与智能体序列 $S^{\mathcal{A}} = (a_1, \ldots, a_T)$。其中，$e_t$ 与 a_t 在时间步 t 上一一对应，用于刻画同步交互过程中的输入输出关系。

2.3.1 联合概率建模视角

从联合建模的角度看，全双工交互过程可以表示为环境序列与智能体序列的联合分布：

$$P(S^{\mathcal{E}}, S^{\mathcal{A}}) = \prod_{t=1}^{T} P\left(e_t, a_t \mid S^{\mathcal{E}}_{<t}, S^{\mathcal{A}}_{<t}\right). \tag{2.1}$$

该建模方式是 NTPP[24] 的基础。此类方法在解码器式 Transformer 中对 (e_t, a_t) 成对进行联合预测，其优化目标可写为：

$$\mathcal{L}_{\text{NTPP}}(\theta) = \mathbb{E}_{(S^{\mathcal{E}}, S^{\mathcal{A}})}\left[\sum_{t=1}^{T} \log P\left(e_t, a_t \mid S^{\mathcal{E}}_{<t}, S^{\mathcal{A}}_{<t}; \theta\right)\right] \tag{2.2}$$

与此相比，早期方法[21] 更多采用条件独立的近似建模方式，通过交叉注意力机制分别预测环境与智能体序列，其优化目标通常是多个条件对数似然的加和，而非严格意义上的联合似然。

2.3.2 条件概率建模视角

对于交互式智能体而言，更常见的建模目标是给定环境输入 $S^{\mathcal{E}}$ 后，预测智能体输出序列 $S^{\mathcal{A}}$，即学习条件分布 $P(S^{\mathcal{A}} \mid S^{\mathcal{E}})$：

$$a_t \sim P\big(a_t \mid S^{\mathcal{E}}_{\leq t},\, S^{\mathcal{A}}_{<t};\, \theta\big)\,. \tag{2.3}$$

在这一设定下，系统在生成当前时刻输出 a_t 的同时，仍需继续接收后续输入 $e_{t+1}, \ldots$，因此模型需要具备编码与解码并行执行的能力[22,30]。同时，全双工对话系统对推理时延具有较高要求，通常需要满足实时交互的约束，即 $\mathrm{Time}(\mathrm{Compute}(a_t)) < 200\,\mathrm{ms}$[41]。此外，输出序列依赖于历史智能体状态 $S^{\mathcal{A}}_{<t}$，这有助于维持生成一致性，并支持回声抑制与上下文保持[26]。

对应的训练目标可表示为：

$$\mathcal{L}_{\mathrm{Cond}}(\theta) = \mathbb{E}_{(S^{\mathcal{E}}, S^{\mathcal{A}})}\left[\sum_{t=1}^{T} \log P\big(a_t \mid S^{\mathcal{E}}_{\leq t},\, S^{\mathcal{A}}_{<t};\, \theta\big)\right] \tag{2.4}$$

在同步数据上采用上述两类目标进行训练，能够使系统在缺乏显式轮次标注的情况下，自发学习到一定的轮替对话模式。

2.3.3 层次化生成与预测式同步机制

以 Moshi[22] 为代表的方法引入中间文本表示 $T^{\mathcal{A}}$，将输出生成过程分解为两个阶段：

$$P(S^{\mathcal{A}} \mid S^{\mathcal{E}}) = \int P\big(S^{\mathcal{A}} \mid T^{\mathcal{A}}, S^{\mathcal{E}}\big)\; P\big(T^{\mathcal{A}} \mid S^{\mathcal{E}}\big)\; dT^{\mathcal{A}} \tag{2.5}$$

该方法先生成文本形式的中间表示，再由文本生成语音，从而在一定程度上结合了文本推理能力与全双工语音交互能力。

SyncLLM[23] 则通过预测用户即将发出的语音片段来降低系统响应延迟，其核心思想可表示为：

$$\hat{e}_{t+1} \sim P\big(\cdot \mid S^{\mathcal{E}}_{\leq t}, S^{\mathcal{A}}_{\leq t}\big)\,, \quad a_{t+1} \sim P\big(\cdot \mid S^{\mathcal{E}}_{\leq t}, \hat{e}_{t+1}, S^{\mathcal{A}}_{\leq t}\big) \tag{2.6}$$

通过对未来输入进行先验预测，系统能够在用户语音尚未完全结束之前提前准备后续输出，从而进一步改善实时交互体验。

2.4 全双工语音语言模型的评测体系

全双工语音语言模型的评测需要覆盖交互时序行为，说话权协调，语义保持和声学表现等多个层面。相较于传统语音系统中以词错误率和主观自然度评分为主的评测方式，全双工场景更加关注系统在持续输入条件下的响应时机、重叠处理、打断恢复和多轮交互稳定性。围绕这些能力，近年来逐步形成了面向全双工交互的专项评测基准与评测流程，包括 Full-Duplex-Bench[39]，Full-Duplex-Bench v1.5[42], Full-Duplex-Bench v2[43], FD-Bench[44], MTR-DuplexBench[45] 和 Talking Turns[46]。

现有评测体系大致可归纳为四类能力维度。第一类是时间动态与响应性评测，主要衡量系统在话轮切换、重叠发声、停止输出、让出话权和响应打断过程中的时间行为，常用指标包括话轮转移偏移、停止时延、让出时延、打断响应时延和流式处理效率。第二类是重叠处理与行为仲裁评测，主要衡量系统对附和、真实打断、话轮切换和非目标干扰的识别与处理能力，相关基准尤其关注重叠语音场景下的行为判别精度与交互稳定性，如 Full-Duplex-Bench v1.5[42]、FD-Bench[44] 和 Talking Turns[46]。第三类是多轮语义一致性与任务完成评测，主要衡量系统在连续全双工交互中的内容相关性、语义连贯性、上下文保持和任务执行能力，这一方向在 Full-Duplex-Bench v2[43] 和 MTR-DuplexBench[45] 中得到了进一步系统化。第四类是声学实现与环境鲁棒性评测，主要关注生成语音的自然度、韵律协调性以及噪声环境、旁人语音和复杂声学条件下的稳定表现。总体来看，这些评测基准共同推动了全双工语音语言模型从单轮低时延响应评估走向面向真实交互过程的系统化评测。

2.5 统一音频表示的基本思想

统一音频表示的核心问题在于刻画不同音频信息在共享潜在空间中的映射关系。其基本思想是通过一个统一的编码机制，将原始音频信号转换为具有结构性的潜在表征，使模型能够在同一表示域中对不同的音频数据进行一致建模。该潜在表征空间承担声学信息压缩功能，使原始音频中高维连续的时序信息转化为更适合学习与推理的结构化表征。

从表征学习的视角来看，音频信号可以被理解为由潜在生成因素共同驱动的观测结果。这些潜在因素可能对应说话人特征，语音内容，情感状态，环境声背景，录制条件等属性。理想的表征应在尽可能保留主要声学信息和语义信息的同时，减少原始观测中

的冗余与噪声。音频表征可以被看作连接原始输入与下游任务之间的中间结构层。在这一过程中，编码器首先从输入音频中提取共享表征，后续模块基于该表示完成具体任务推断。从时间结构的角度看，音频表征具有明显的序列属性。实际建模中，统一音频表征同时包含两类互补的信息形态，面向时间序列的局部表示和面向整段样本的全局表示和从整体上概括音频的主要属性与高层语义。统一音频表示的基本形式，可以理解为在共享编码器的基础上，同时学习局部结构表征与全局结构表征，并通过一致性的训练机制，使不同粒度上的表征在几何性质与统计性质上保持协调。

3 低延迟响应与自然交互的全双工对话系统设计与实现

3.1 系统设计目标与总体思路

本章重点讨论系统如何在现实语音交互场景中同时满足两个核心目标。系统能够在用户表达趋于完成时及时接管话轮，降低等待时间的低延迟响应目标和系统能够处理打断，附和，重复发言，短暂停顿和背景第三方语音等复杂现象的自然交互目标。

基于当前资源条件和工程实现的实际需求，我们采用相较完全端到端的模型更为灵活的半级联方案。我们的系统建立在已有语音活动检测模型（VAD）、自动语音识别模型（ASR）、多模态大语言模型（MLLM）和语音合成模型（TTS）之上，通过模块内部的状态管理和针对特定任务的优化，解决现实交互中常见的问题。该方案既保留了VAD、ASR、TTS 等模块化系统的可部署性，又通过多模态大语言模型直接读取用户音频，增强了对副语言线索和复杂交互状态的处理能力。我们认为一个面向实际应用的全双工系统应在响应速度和行为合理性之间达成平衡。如果系统虽然开口很快，但频繁误打断用户，或者在被打断后无法正确恢复，那么其交互质量仍然难以令人满意。所以我们的主要目标是构建一个系统，主要围绕状态管理这一核心机制，把不同模块组织成一个可持续响应与动态交换说话权的整体。

3.2 需求分析与关键问题

一个可靠的全双工系统应该具备以下能力。第一，系统必须能够判断用户是否已经完成当前表达。在自然对话中，用户常常会出现短暂停顿，迟疑词或未完成句，若系统仅依据短时静音给出应答，就会造成不自然抢话。第二，系统必须能够区分附和与真实打断。用户在系统播报期间发出的“嗯”“对”“好的”等短反馈，通常表示继续关注而非争夺说话权。若系统将其一律视作打断，则会频繁中止自身回复。第三，系统需要具备面向复杂环境的鲁棒性，例如背景中非目标说话者发声或环境声音都可能干扰系统状态判断。第四，系统还需要在尽量低的推理开销下实现稳定运行，从而支撑真正的实时交互。我们采用对话单元加状态切换的总体思路，将复杂交互行为转化为模型可持续处理的局部决策问题。

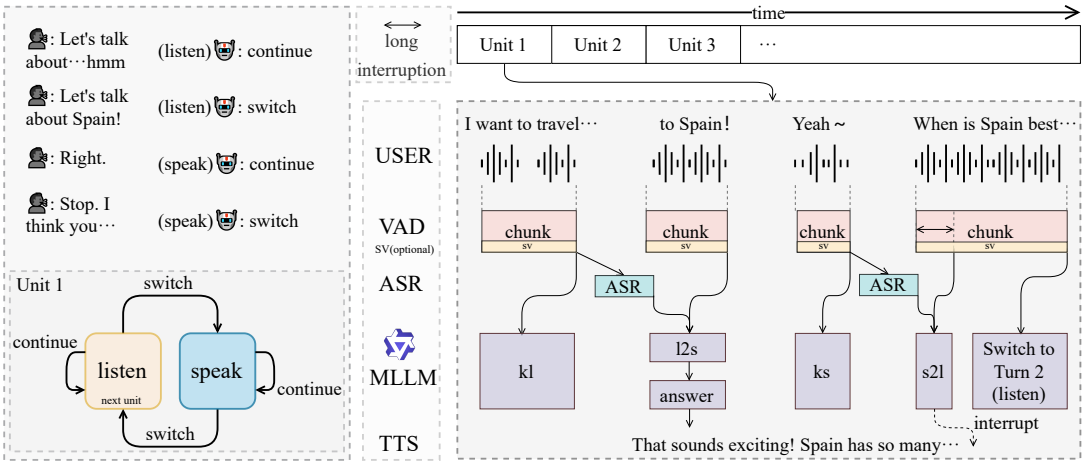

图 3.1 基于单元的全双工对话概览。左上展示了 MLLM 的控制行为示例；左下展示了由 continue/switch 信号驱动的单元内听/说状态切换。右侧展示了单个单元的执行过程，其中，kl（保持聆听）和 l2s（由听转说）对应于聆听状态下的 continue/switch，而 ks（保持说话）和 s2l（由说转听）对应于说话状态下的 continue/switch。

3.3 基于对话单元的状态转换机制

我们的系统将自然对话划分为一系列最小对话单元，每个单元内部仅包含监听状态和说话状态两种基本状态。系统从监听状态开始，在多模态大语言模型的控制下，通过 continue 与 switch 两类动作实现状态保持或状态切换。当系统处于监听状态时，continue 表示继续接收用户语音，switch 表示用户当前发言已具备足够的语义完整性，系统可以切换到说话状态并开始回复；当系统处于说话状态时，continue 表示系统继续播报当前回答，switch 表示用户发声构成真实打断，系统应立即停止播报并回到监听状态。我们将复杂的全双工交互问题转化为一个统一的状态决策问题。与传统规则驱动系统相比，系统需要模型结合声学和语义和上下文信息综合判断此刻是否应当维持当前状态。例如，在监听状态下，短暂停顿并不一定意味着用户说完，系统还需要结合语义完整度和语气走势判断是否应当继续等待。在说话状态下，短反馈也并不必然意味着用户要夺回说话权，系统需要进一步区分附和与真实打断。

由于每一次决策都可以明确归类为某一状态下的保持或切换，我们的系统的行为轨迹更容易被追踪和分析。相比设计大量细碎的交互标签或复杂层级状态，二元动作空间更适合作为多模态大语言模型的控制输出接口，也更方便与下游 TTS 中断、ASR 缓存和前端流式通信机制对接。

3.4 系统总体架构与模块功能

在总体结构上，系统由音频采集模块，上下文模块，决策模块和语音生成模块四部分组成，各模块在状态转换机制的统一控制下协同工作。

音频采集模块负责持续接收用户语音流，并通过语音活动检测判断当前是否存在有效语音输入。在需要区分说话人身份的场景中，还可引入说话人验证组件，以过滤非目标说话者或背景第三方语音。从功能定位上看，该模块是系统的前端感知层，它为后续模块提供低延迟的语音起止信息和说话人相关线索，是实现实时交互的前提。

上下文模块采用异步 ASR 进行实时转写，但与传统级联系统不同，这里的 ASR 文本不直接决定系统回复，而是作为辅助语义信息缓存到上下文中。也就是说，文本转写在本系统中被视为增强信息而不是唯一语义通路。这样做的核心目的是避免系统控制逻辑完全依赖延迟较高且可能存在识别误差的转写结果，从而缓解传统级联方案在复杂交互条件下的问题。

决策模块是整个系统的核心，由多模态大语言模型承担。我们采用 Qwen3-Omni[47] 作为多模态决策模型，使其直接接收用户音频和缓存文本上下文，并在当前状态条件下输出 continue 或 switch 决策。该模块承担了三类职责：第一，理解用户当前输入的内容和表达状态。第二，判断说话权是否应当发生转移。第三，在系统回复期间判断新语音是附和还是打断。相较于传统由 ASR 输出进入 LLM 的文本级串行链路，决策模块能够更早地使用声学与副语言信息，因此更适合服务于全双工状态控制。

语音生成模块负责使用 TTS 合成系统回复，并在播报期间持续接受来自决策模块的控制信号。一旦系统在 speak 状态下判定用户发声构成真实打断，TTS 播报会立即中止，系统重新转入 listen 状态。语音生成模块是需要被实时调度，可被中断，并与状态机严格耦合的交互模块。

3.5 系统实现与部署方式

在系统实现层面，我们采用开源组件完成整体搭建，各功能模块配置如下。语音活动检测模块采用 Silero VAD[48]，说话人验证模块采用 CAM++[49]，异步自动语音识别模块采用 Paraformer[50]，多模态决策模型采用 Qwen3-Omni[47]，语音合成模块采用 IndexTTS 1.5[51]。实验运行环境配置为两张 NVIDIA A100 GPU。为满足实时推理需求，Qwen3-Omni[47] 与 IndexTTS 1.5[51] 均基于 vLLM 进行部署。系统前后端通信采用 WebSocket 与

FastAPI 实现。前端可持续发送用户音频流，后端在检测到状态变化或生成回复后，实时返回控制结果与语音输出。与完全离线处理方式相比，该实现方案更接近真实语音对话系统的运行模式，因此能够更准确地反映系统在实际应用场景下的响应延迟与交互性能。

3.6 评测方案与指标设置

我们主要采用 Full-Duplex-Bench v1.5[39]及相关挑战赛数据作为主要评测参照。该类评测尤其关注停止延迟、响应延迟、恢复/拒识表现以及交互行为的合理性，能够较好对应本文所关心的系统目标。

从指标意义上看，首帧响应延迟用于衡量系统在监听状态下接管话轮的速度，停止或总延迟用于衡量系统在被打断或状态切换时的响应效率，打断场景相关分数用于衡量系统对真实打断行为的识别与处理能力，恢复或拒答相关分数则能够反映系统在复杂交互状态下是否具备较好的行为稳定性。相比仅使用识别准确率或语音自然度，这些指标更能体现全双工对话系统的真实交互质量。

表 3.1 全双工对话系统在相关数据集上的实验结果

模型	打断场景（5 个）			拒答场景（4 个）				总体指标			
	平均响应分数↑	平均停止时延↓	平均响应时延↓	用户实时附和恢复分数↑	第三方语音后恢复分数↑	暂停处理拒答率↑	向第三方说话恢复分数↑	首帧响应时延↓	打断总分↑	拒答总分↑	总时延↓
Freeze-Omni[52]	0.591	1.762s	2.039s	0.640	0.291	0.425	0.160	1.721s	0.591	0.379	1.841s
Moshi[22]	0.635	1.359s	3.672s	0.500	0.190	0.208	0.140	3.184s	0.635	0.230	2.738s
OSUM-EChat[53]	0.802	1.871s	2.685s	0.810	0.110	0.853	0.050	2.753s	0.802	0.459	2.436s
ours（验证集）	0.897	1.106s	2.461s	0.765	0.340	0.830	0.235	1.528s	0.897	0.500	1.698s
ours（测试集）	0.897	1.632s	-	-	-	-	-	1.632s	0.897	0.578	1.632s

注：下划线表示该列最优结果；↑表示数值越大越好，↓表示数值越小越好。

3.7 实验结果与分析

从整体结果看，系统在响应速度和交互行为两个方面均优于基线系统。首先，在首响应延迟方面，系统从 2.753s 降低到验证集的 1.528s，说明其在用户发言结束附近具备更快的接话能力。造成这一提升的主要原因，是系统不再完全等待稳定文本转写结果才进入回复阶段，而是通过多模态大语言模型更早地对当前音频状态进行判断，从而缩短了传统级联管线中的累计等待时间。

在打断相关得分上，系统取得了 0.897 的结果，说明基于对话单元的状态建模在区分真实打断与非打断发声方面具备明显优势。传统轮次式语音系统往往只能在用户停止之后再进入回复，或者在听到新语音后机械地停止播报，缺乏对附和，短反馈和持续打断之间差异的理解。系统通过把这些现象纳入统一状态控制框架中进行判断，提高了交互行为的合理性。在恢复或拒识相关总分上，测试集结果达到 0.578，表明系统在复杂状态切换条件下具有更好的稳定性。最后，在总延迟指标上，系统由基线的 2.436s 降至 1.632s，进一步说明半级联方案在整体响应链路上具有明显优势。

如果进一步结合系统结构进行分析，可以发现改进并非来源于单一模块替换，而是来源于整体交互链路的重新组织。本文系统将音频感知，状态判断和回复生成三者更紧密地连接起来，使模型在监听阶段就开始积累话轮决策所需信息，在说话阶段又持续监控新的输入变化，从而形成提前感知，持续监控，即时判断的处理模式。这一模式使系统能够同时向低延迟和自然交互两个方向优化，而不是单纯牺牲一方去换取另一方。

3.8 本章小结

我们构建了一个以低延迟响应与自然交互为目标的半级联全双工对话系统。系统以状态管理为核心，通过 VAD、ASR、多模态大语言模型与 TTS 的协同工作，实现了对打断，附和和话轮切换等复杂行为的统一处理。实验结果表明，该方法在响应速度和交互合理性两方面均优于基线方案，体现了我们的系统目标。

4 音频理解模块探索与实验分析

4.1 研究目的

第三章面向全双工语音交互任务，构建了以 Qwen3-Omni[47] 为核心的多模态决策系统，并验证了其在低延迟响应，自然打断处理和连续交互控制中的有效性。Qwen3-Omni[47] 作为大规模多模态基础模型，依托文本，音频与视频等多模态联合预训练，具备较强的通用音频理解与决策能力，为系统的状态建模提供了坚实基础。然而针对语音交互场景，系统的核心输入主要仍然集中于语音及其伴随的环境声，说话人状态和副语言线索，视频模态在纯语音对话场景下并非必要信息。对于以实时性和自然交互为目标的系统而言，直接使用大规模的全模态预训练 LLM，虽然能够带来较强的通用能力，但同样造成了模态冗余。

因此我们针对音频理解模块进行改进，重新审视在交互式语音系统中何种音频表征方式更适合作为决策前端。一方面，交互系统要求前端表示能够保留语音内容，说话人属性，情绪状态和时序细节等信息，以支撑轮次判断和语义理解等任务；另一方面，真实场景中的输入并不局限于纯净语音，还包含背景噪声，环境事件以及复杂非语音声音，这又要求表征具备更广泛的跨域音频建模能力。现有音频编码器在这两类需求之间往往存在明显失衡，面向语音任务优化的模型通常具备较强的语音识别与细粒度时序建模能力，但对复杂非语音场景的语义覆盖不足。面向通用音频的模型虽然能够覆盖环境声与音乐等更广泛的声音类型，却往往难以兼顾语音任务所需的精细结构信息。已有工作通过多编码器融合扩展覆盖范围，但这类方案通常伴随更高的结构复杂度，额外的表示对齐成本以及音频 token 冗余，与低延迟交互系统对高效性的要求并不完全一致。因此，我们引入 UniWhisper 所代表的统一音频表示思路，尝试从单编码器，统一监督和持续多任务训练的角度重新设计音频理解模块。该类方法的核心特点在于以单一音频编码器作为统一表征骨架，将语音，环境声和音乐等异构任务统一映射到一致的指令—回答训练范式中，通过共享监督接口提升编码器的跨域泛化能力。在避免多编码器特征拼接和额外对齐开销的同时，尽可能保留语音任务所需的细粒度声学线索，并增强对非语音语义信息的建模能力。我们的的研究目的在于探索对于以低延迟响应和自然交互为目标的语音对话系统，是否能够以一种更精简，更统一且更稳健的音频编码路线，在维持通用理解能力的同时减少模态与结构冗余，从而为后续多模态决策提供更高质量，更适

合实时交互场景的输入表征。

4.2 方案概述

Whisper 采用编码器—解码器结构的 Transformer 模型。给定输入音频 x，首先提取其对数梅尔频谱特征，并将其编码为声学表示；随后，解码器在音频表示及先前已生成 token 的条件下预测下一个文本 token，训练目标为标准的交叉熵损失。

UniWhisper 在保留下一个 token 预测这一标准训练目标的基础上，对解码器结构进行了调整，以更好地适应指令式监督学习范式。具体而言，其编码器由 Whisper Large-v3[36] 初始化，自回归解码器则采用预训练语言模型 Qwen3-0.6B[6]。同时，引入一个轻量级适配器，用于将编码器输出的隐状态映射到解码器所需的隐藏维度。

设 $\mathrm{H} \in \mathbb{R}^{T' \times d_w}$ 表示 Whisper 编码器的输出，其中 T' 为编码后的时间步数，d_w 为编码器隐藏维度；设 d_q 为解码器的隐藏维度。适配器的输出定义为：

$$\mathrm{Z} = A(\mathrm{H}) \in \mathbb{R}^{T' \times d_q}, \tag{4.1}$$

其中，$A(\cdot)$ 为一个小型多层感知机（MLP），用于实现从 d_w 到 d_q 的可学习投影。

将预训练语言模型作为解码器，能够为模型提供更强的语言先验，从而更好地匹配指令跟随式任务的目标形式，并在一定程度上缓解音频表示与文本语义之间的对齐难题。在训练过程中，预训练语言模型解码器始终保持冻结，仅更新编码器和投影模块。

UniWhisper 将多种音频任务统一表示为指令—回答的形式，涵盖自动语音识别（ASR），语音翻译，音频描述，关键词或属性预测，音频问答以及音频文本匹配等任务。如图4.1 所示，每个训练样本均表示为一个 prompt，其中将控制标签与自然语言指令相结合。控制标签可以包括任务类型，音频语言，文本语言，可选时间戳以及输出约束等信息，而监督信号则以文本回答的形式提供。因此，不同任务之间的差异主要体现在 prompt 内容和目标答案上，而训练与解码接口则保持统一。

给定音频前缀表示 Z，prompt token u 以及回答 token 序列 $\mathrm{v} = (v_1, \ldots, v_{|\mathrm{v}|})$，解码器建模为：

$$p_\theta(v_t \mid v_{<t}, \mathrm{u}, \mathrm{Z}), \tag{4.2}$$

并在回答 token 上优化下一个 token 预测的交叉熵损失：

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{t=1}^{|\mathrm{v}|} \log p_\theta(v_t \mid v_{<t}, \mathrm{u}, \mathrm{Z}). \tag{4.3}$$

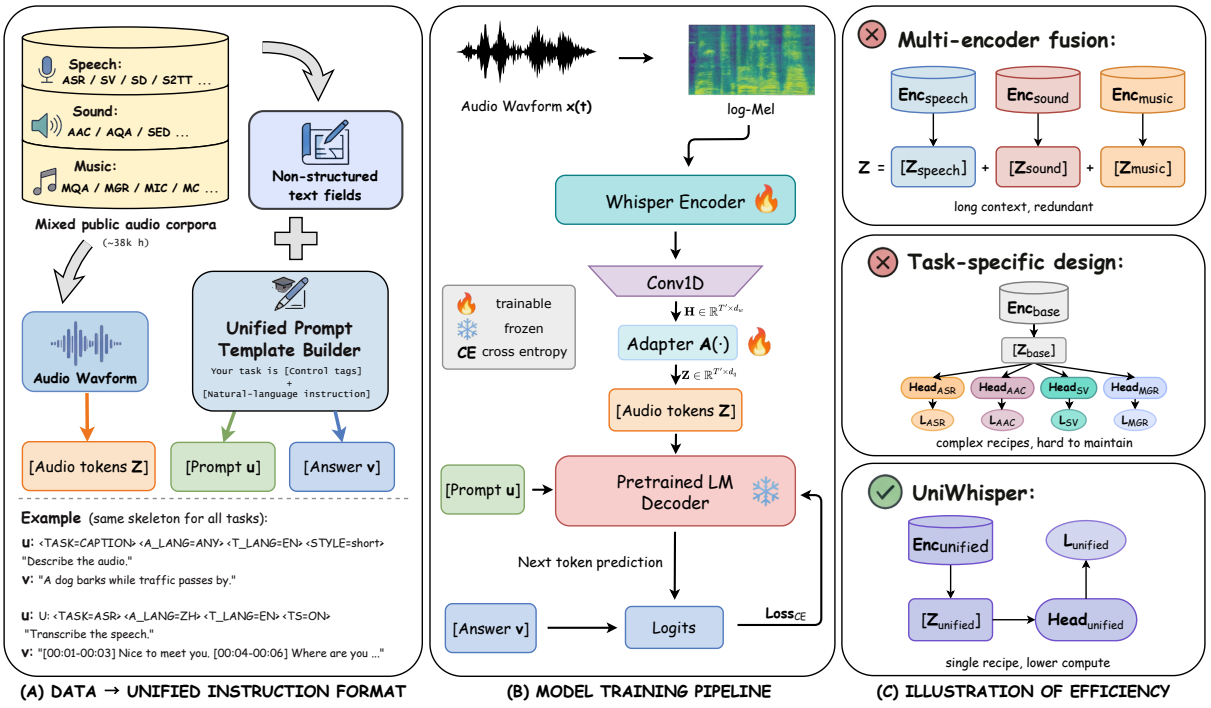

图 4.1 UniWhisper 持续多任务训练框架概览。(a) 将异构数据集转换为统一的指令—回答格式。(b) 使用单一音频编码器，并在回答 token 上进行下一个 token 预测训练。(c) 与常见替代方案的比较，突出我们在减少音频 token 冗余和统一监督接口方面的优势。

由于 UniWhisper 仅使用单一音频 token 流 Z 作为前缀，因此避免了多编码器特征拼接所带来的音频 token 冗余问题。其多领域能力主要来源于统一的任务模板，以及在持续多任务训练框架下对混合多任务数据的联合学习。

4.3 实验设置

4.3.1 数据集

我们使用一组开源音频数据集对 UniWhisper 进行训练，具体包括 General，Speech，Sound 和 Music 四类。为了保持一致性，我们在将所有数据集转换为统一的 instruction–answer 格式时保留了原始监督信号。为避免训练集与评测集之间的信息泄漏，我们采用基于标识符的过滤，并结合哈希与声学指纹进行内容级去重。训练数据包括 AudioCaps[54]（音频描述生成，总计 142.5 小时），AudioSet[55]（音频标签分类，总计 5.8k 小时），LAION-Audio[56]（音频描述生成，总计 4.3k 小时），WavCaps[57]（音频描述生成，总计 7.6k 小时），AISHELL-1[58]（语音识别与说话人识别，总计 178 小时），GigaSpeech[59]（语音识别，总计 10k 小时），Libri-adhoc40[60]（语音识别，总计 4.5k 小时），LibriSpeech[61]（语音

识别与说话人验证，总计 860 小时），Clotho[62]（音频描述生成，总计 43.6 小时），Seeing Sound[63]（声音事件检测，总计 0.2 小时），SONYC-UST[64]（音频标签分类，总计 51.5 小时），TAU-ASC2020[65]（声学场景分类，总计 64 小时），URBAN-SED[66]（声音事件检测，总计 27.8 小时），VGGSound[67]（声学场景分类，总计 553.3 小时），GuitarSet[68]（吉他转录，总计 3 小时），MAESTRO[69]（钢琴转录，总计 200 小时），MedleyDB[70]（自动音乐转录，总计 7.3 小时），MTG-Jamendo[71]（音频标签分类，总计 3.8k 小时），MusicCaps[72]（音乐描述生成，总计 15.3 小时），MusicNet[73]（自动音乐转录，总计 34 小时），Slakh2100[74]（自动音乐转录，总计 145 小时），SongDescriber[75]（音乐描述生成，总计 23 小时）以及 YT8M-MTC[76]（音乐描述生成，总计 11.7 小时）。

评测集基于 HEAREval[77] 构建，并在涵盖对话，环境声音和音乐的 20 个任务上进行评测。由于 HEAR 对 human voice processing 的覆盖有限[78]，我们进一步补充了若干面向对话的任务。更多细节可参考 HEAR[77] 与 X-ARES[78]。

4.3.2 训练细节

我们将所有音频重采样至 16 kHz，并使用 25 ms 窗口和 10 ms 帧移提取 128 维对数梅尔频谱特征。训练时统一使用 30 s 音频片段，较短片段采用零填充。为降低音频序列长度，我们额外采用时间步幅为 2 的时间下采样卷积。结合 Whisper 编码器自身的下采样后，每个输出帧大约对应输入波形的 40 ms。我们在编码器中传递填充掩码，以确保填充帧不参与注意力计算。训练时仅在回答 token 上计算下一 token 预测的交叉熵损失，并对 prompt token 进行掩码处理。优化器采用 8 位 AdamW，并使用余弦衰减学习率策略，初始学习率设为 2×10^{-5}，权重衰减设为 0.01。训练前进行 1,500 步预热，以对齐音频与文本表示空间，总训练步数为 30,000。训练采用 bf16 和分布式数据并行，在 8 张 A800 GPU 上完成，总实际训练时间约为 24 小时，每卡批大小为 32，对应全局批大小 256。除非另有说明，我们仅更新 Whisper 编码器和投影适配器，冻结预训练的语言模型解码器。

4.3.3 评测协议

我们使用两种协议评估编码器表征，即基于浅层 MLP 的监督探测和非参数的 kNN。任务分为片段级任务和帧级任务。前者为每个样本使用一个表征向量，后者使用一段表征序列。我们还在 LibriSpeech-100h[61] 子集上加入了自动语音识别任务。

表 4.1 MLP 和 kNN 在 20 个任务上的逐任务归一化结果

领域	任务	wav2vec 2.0		HuBERT		WavLM		BEATs		CLAP		Whisper		CLAP-U		Whisper-U		Whisper-U-3		ours	
		MLP	kNN	MLP	kNN	MLP	kNN	MLP	kNN	MLP	kNN	MLP	kNN	MLP	kNN	MLP	kNN	MLP	kNN	MLP	kNN
Speech	ASV2015	94.0	95.8	95.1	88.5	97.4	94.6	91.5	81.0	95.1	71.2	97.5	90.3	92.3	82.9	98.3	95.0	94.7	90.6	99.0	92.9
	CREMA-D	55.5	21.8	64.5	34.0	71.3	28.0	66.5	39.7	36.9	25.4	57.5	38.2	54.3	49.0	64.3	17.8	66.3	30.0	84.4	67.9
	FSC	47.2	1.6	96.5	3.6	97.5	4.3	47.3	5.5	3.6	1.6	97.8	23.4	17.7	5.4	97.2	1.0	98.0	53.1	82.7	14.6
	LibriCount	58.0	20.8	58.2	18.2	65.1	38.8	68.2	37.8	38.8	28.1	59.4	45.2	60.9	54.0	62.8	9.1	53.7	41.4	64.4	47.2
	LS-100h	16.0	–	83.4	–	64.1	–	15.8	–	0.2	–	62.5	–	1.5	–	79.0	–	89.9	–	91.6	–
	LS-MF	95.3	70.3	97.8	79.9	97.9	77.2	97.5	92.3	90.2	86.8	90.8	58.0	97.1	96.4	97.2	53.9	82.7	55.9	98.1	91.3
	RAVDESS	44.1	15.8	58.1	32.9	68.0	25.2	66.6	37.9	23.7	21.9	61.3	38.5	55.9	45.5	59.4	23.9	65.6	36.8	88.2	70.0
	GSC	71.2	23.2	95.7	28.9	96.3	58.3	88.0	49.6	25.1	13.7	97.0	74.8	67.0	49.1	96.7	9.8	94.9	73.0	95.5	65.4
	VocalSound	77.3	24.1	85.5	35.7	89.3	31.1	91.4	75.2	42.6	25.9	89.3	57.1	88.5	85.8	90.7	39.0	90.6	43.3	93.0	90.9
	VoxCeleb1	34.0	0.2	58.2	6.5	65.4	8.9	41.3	13.2	4.3	0.5	22.0	4.1	13.6	7.8	40.9	0.9	33.8	4.1	45.5	4.5
	VoxLingua33	55.8	0.6	75.0	6.3	86.2	29.0	41.9	14.8	7.7	3.5	97.6	95.8	18.8	13.4	97.6	53.6	94.0	62.8	89.5	71.6
Music	FMA	47.8	21.1	50.5	22.6	52.6	31.0	66.0	61.0	30.0	25.7	57.4	51.4	66.0	62.2	61.0	40.8	61.2	46.5	68.9	67.3
	GTZAN	63.7	31.1	69.6	20.7	74.0	43.6	89.8	85.2	40.4	34.9	71.2	53.3	84.2	79.5	70.5	11.2	82.3	64.1	94.5	89.5
	NSynth	31.6	25.1	37.7	35.5	40.9	25.0	64.8	61.6	24.5	20.1	46.0	12.4	77.3	78.0	47.0	15.0	49.7	38.5	70.7	70.0
Sound	DESED	31.7	–	33.5	–	38.8	–	4.2	–	2.4	–	29.4	–	20.2	0.0	31.0	–	44.3	–	57.0	–
	ESC-50	51.9	7.9	56.9	13.2	65.7	19.2	95.3	85.5	16.3	18.2	62.4	29.5	97.0	95.6	67.0	3.5	80.6	46.4	95.8	93.7
	FSD18-Kaggle	23.1	–	34.9	–	36.4	–	79.1	–	6.4	–	36.0	–	87.6	0.0	21.8	–	55.8	–	90.5	–
	FSD50k	16.5	–	21.9	–	27.4	–	56.7	–	4.5	–	32.1	–	59.1	0.0	34.6	–	50.3	–	60.0	–
	UrbanSound 8k	66.7	34.8	65.8	34.3	69.1	30.6	89.3	81.4	36.4	33.8	71.9	45.5	85.4	83.1	73.4	11.5	76.4	54.4	84.4	78.3
	Vocal Imitation	14.5	0.9	17.4	2.4	24.5	4.3	24.1	13.4	2.5	1.7	24.1	5.2	17.2	11.4	26.8	1.6	28.5	10.9	25.7	14.7
Weighted Avg.		42.8	27.7	69.2	32.6	67.2	37.1	52.4	49.2	22.0	27.6	64.2	45.7	43.9	53.1	70.3	27.8	74.4	47.0	80.9	60.4

注：各任务得分根据任务特定的上下界归一化到 $[0, 1]$ 区间，数值越大表示性能越好，并以百分比形式展示。下划线表示该列最优结果。

具体而言，我们使用编码器最后一层的帧级表征作为输入。对于片段级任务，我们沿时间维进行平均池化，得到单个片段表征；对于帧级任务，则保留完整的帧序列，并在需要时通过填充对齐帧标签。对于 MLP 评测，我们冻结编码器，并在每个任务上使用浅层 MLP 对片段表征或帧序列进行监督训练；对于 kNN 评测，我们直接在池化后的表征空间中进行分类或检索，而不训练额外的分类器。

评价指标方面，我们使用最小—最大归一化将每个指标映射到 $[0, 1]$，并用任务可达的最好值与最差值进行归一化，其中 Acc，mAP，Seg-F1 和 Recall@1 分别使用 1 和 0 作为上下界；对于 ASR，我们使用 $\text{iWER} = \max(1 - \text{WER}, 0)$。随后计算 $S = \frac{\sum_i n_i \hat{M}_i}{\sum_i n_i}$，并在表4.1中报告 MLP 和 kNN 的结果。

从结果特点看，UniWhisper 在非语音任务上提升较为明显，同时仍然保持了较强的语音相关能力。这说明它并不是通过放弃语音性能来换取一般音频语义，而是在多个领

域之间获得了较好的平衡。对于全双工系统而言，这种平衡能力恰恰具有潜在价值，因为真实对话场景中并不只有干净语音输入，系统往往还需要面对复杂背景声和多种非语言音频线索。

4.4 实验结果

我们在相同评测协议下，对若干广泛使用的预训练音频编码器进行了比较，包括 wav2vec 2.0-Large[79]，HuBERT-Large[34]，WavLM-Large[35]，Whisper-Large-v3[36]，BEATs-iter3[80] 和 CLAP-HTSAT[81]。各任务的详细结果见表4.1

4.4.1 MLP 结果

表 4.1 表明，UniWhisper 在 MLP 和 kNN 评测下的加权平均分分别达到 80.9 和 60.4。这一提升说明，在统一指令监督下进行持续多任务训练，可以增加表征中能够被线性访问的信息量。我们观察到，最大的提升主要出现在依赖全局语义线索的非语音任务上，例如音频标签和音频描述生成任务。此外，UniWhisper 在面向语音的任务上也保持了较强性能，包括说话人分类和副语言分类。与领域专用基线相比，UniWhisper 在环境声音和音乐任务上缩小了性能差距，同时不需要多编码器特征融合。

我们还注意到，wav2vec[79] 和 WavLM[35] 这类面向语音的编码器在语音占比较高的任务子集上仍然表现强劲，但在统一评测设置下，它们在音乐和复杂声景任务上的表现一致性较弱。相对地，CLAP[81] 和 BEATs[80] 在与其预训练目标一致的任务上表现较好，但在依赖语音细粒度语音学信息的任务上往往较弱。UniWhisper 以 Whisper[36] 的声学表征为基础，并加入显式针对非语音语义的监督信号，从而改善了跨领域性能的平衡性。

4.4.2 kNN 结果

表 4.1 中的 kNN 结果整体上与 MLP 的趋势一致，但更强调嵌入向量空间的几何结构，因为这一评测不引入任务特定的预测头。UniWhisper 在 kNN 上相较 Whisper[36] 也有提升，这表明基于指令的持续训练不仅提高了探测准确率，也使表征空间的组织更加合理。在各个基线中我们发现，像 CLAP[81] 这样针对全局对齐优化的模型在面向检索的任务上可以表现较好，但在细粒度语音迁移任务上仍然较弱。这一现象与检索导向编码器的特性一致，即其表征更偏向全局语义对齐，而非稠密的时序细节。

4.4.3 消融实验

我们在相同的指令式持续训练流程下，对编码器选择和解码器设计进行了消融分析。受 CLIP[82] 的启发，我们测试了一种 CLIP 风格的对比学习编码器，即在保持其余训练流程不变的情况下，用 CLAP-HASAT[83] 替换我们的编码器。为保证公平，CLAP-Uni 绕过了 CLAP 最终的片段级平均池化，转而使用倒数第二层的稠密特征；对于每个 10 s 音频片段，得到 $h \in \mathbb{R}^{64\times 2048}$，其时间步长约为 156 ms。我们同时移除了 UniWhisper 中额外使用的时间卷积层。

如表4.1 所示，我们的指令式持续训练显著提升了 CLAP 的表现：CLAP-Uni 的加权平均分几乎翻倍，并且在与 CLAP 预训练目标一致的语义音频任务上可与 UniWhisper 竞争。然而，它在需要精确时序建模的语音理解任务上的提升不如前者明显。这说明，即使经过持续训练，骨干网络原生的时间粒度仍然是影响性能的关键因素。我们还比较了预训练语言模型解码器与原始 Whisper 解码器在相同流程下的表现。如表4.1 所示，Whisper 解码器的增益更慢，语义对齐能力也更弱：Whisper-Uni-1 在 MLP 上达到 0.70，但在 kNN 上下降到 0.28，甚至低于原始 Whisper 的 kNN 分数 0.46。这表明，尽管其 probe accuracy 更高，但其嵌入向量空间的结构性更差。多轮 replay 可以进一步改善结果，但代价显著更高，且最终仍低于 UniWhisper。这一现象支持采用预训练语言模型解码器，因为在固定计算预算下，它更有利于对齐指令式语义，并保持表征结构。

4.5 本章小结

我们从音频理解子模块增强的角度，对 UniWhisper 所代表的统一音频表示方法进行了补充分析。相关结果表明，该类方法能够在语音，环境声和音乐多个领域之间形成更平衡的音频表示能力。这一部分的意义主要在于为第三章主体系统提供后续扩展方向，即在保持全双工系统总体框架不变的前提下，可以通过增强音频理解模块来进一步提升系统在复杂场景中的自然交互能力与鲁棒性。

5 综合分析与讨论

5.1 当前研究仍存在的主要问题

尽管相关研究已取得明显进展，但全双工对话系统领域仍面临很多共性挑战。

数据问题仍然突出。真正高质量的同步多通道自然对话数据十分稀缺，尤其是中文，多语种和真实复杂场景数据更加不足。对于全双工对话模型而言，数据不仅要覆盖广泛的对话内容，还需要覆盖何时说，如何打断，如何附和，如何恢复等时序行为。对于统一音频表示方法而言，数据又必须同时覆盖语音，环境声和音乐等多个领域。现实中，这两类需求很难在同一数据体系中同时满足。因此，数据瓶颈会同时制约全双工对话学习与统一音频表示训练。

评价体系尚需进一步统一。全双工交互系统涉及多类指标，不同工作之间的实验设置与评测标准仍存在差异。统一音频表示研究虽然在跨任务评测方面较为完整，但与真实交互场景的结合仍有进一步提升空间。例如，一个在离线音频分类基准上表现优秀的编码器，并不必然能在全双工对话场景中稳定支持说话权切换判断。同样一个在小规模交互数据集上表现良好的全双工系统，也未必能在复杂开放环境下保持鲁棒性。因此，如何把离线表征评测与在线交互评测更有效地结合起来，是未来需要重点解决的问题。

第三，体系结构仍缺乏真正意义上的统一。当前全双工系统中既有工程化同步路线，也有学习式同步路线。统一音频表示中既有单编码器持续训练方法，也有多编码器融合方法。不同路线各有优势，但尚未形成被广泛接受的主流范式。这种分化在早期研究阶段是正常现象，但也意味着系统间难以直接迁移经验，研究积累容易碎片化。

第四，系统部署与安全控制仍然困难。低时延全双工系统虽然更接近自然交互，但如果缺乏充分的安全过滤，上下文管理和异常恢复机制，系统可能在误触发，误打断和不适当生成方面面临更高风险。特别是在系统具备主动插话和边听边说能力后，错误行为可能比传统轮次式系统更频繁地暴露给用户，因此安全机制不能再作为部署后的附属模块，而应当在模型设计与系统评测阶段同步考虑。

5.2 全文总结

我们围绕低延迟响应与自然交互的全双工对话系统研究进行了系统整理与分析。首先，通过对全双工语音语言模型综述论文的归纳，明确了全双工交互从轮次式向同步式

发展的技术趋势，并总结了工程化同步与学习式同步两类主要方法，以及覆盖时序行为，语义一致性和声学质量的多维评测框架。我们认为真正的全双工能力并不等价于简单的低延迟，而是涉及输入感知，输出生成，行为判断和系统响应之间的认知并行性。其次，围绕半级联全双工语音对话系统，我们分析了基于对话单元的状态切换机制及其工程实现方式，说明了多模态大语言模型在低时延、可打断语音交互中的应用潜力。该方法通过把复杂对话流拆解为局部单元，并把 continue/switch 作为统一决策动作，在保持工程可部署性的同时显著改善了系统的交互灵活性。最后，围绕 UniWhisper 所代表的统一音频表示框架，我们从音频理解子模块增强的角度总结了其统一指令监督和持续多任务训练思路，并讨论了其对全双工系统复杂场景感知能力提升的潜在意义。

总体而言，我们认为更自然的全双工对话系统需要同时具备两个层面的能力。其一，是在系统主线层面实现边听边说，及时打断和动态轮换的全双工对话能力。其二，具备更稳健的音频理解能力，为状态判断和复杂场景鲁棒性提供支撑。前者决定系统是否能够实现低延迟响应，后者决定系统在复杂现实环境中的自然交互。将系统设计与音频理解增强有效结合是进一步推动全双工对话系统从可用走向自然与鲁棒的重要方向。

参 考 文 献

[1] OPENAI, et al. GPT-4o system card[A]. 2024. arXiv: 2410.21276.

[2] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[3] GUO D, YANG D, ZHANG H, et al. Deepseek-R1: Incentivizing reasoning capability in llms via reinforcement learning[A]. 2025.

[4] SAKSHI S, TYAGI U, KUMAR S, et al. Mmau: A massive multi-task audio understanding and reasoning benchmark[A]. 2024.

[5] BAI Y, CHEN J, CHEN J, et al. Seed-ASR: Understanding diverse speech and contexts with llm-based speech recognition[A]. 2024.

[6] Qwen Team. Qwen3 technical report[A/OL]. 2025. https://arxiv.org/abs/2505.09388. DOI: 10.48550/arXiv.2505.09388.

[7] HU H, ZHU X, HE T, et al. Qwen3-tts technical report[A]. 2026.

[8] ARORA S, CHANG K W, CHIEN C M, et al. On the landscape of spoken language models: A comprehensive survey[A]. 2025.

[9] LIAO Z, et al. Flexduo: A pluggable system for enhancing spoken dialogue models with full-duplex capabilities[A]. 2025. arXiv: 2502.08763.

[10] SHI M, et al. Semantic VAD: Low-latency voice activity detection for speech interaction [A]. 2023. arXiv: 2305.12450.

[11] WU W, GUAN W, WANG K, et al. Phoenix-vad: Streaming semantic endpoint detection for full-duplex speech interaction[A]. 2025.

[12] YAN R, CHEN W, LIU Z, et al. Soulx-duplug: Plug-and-play streaming state prediction module for realtime full-duplex speech conversation[A]. 2026.

[13] FU C, et al. Vita-1.5: Towards GPT-4o level real-time vision and speech interaction[A]. 2025. arXiv: 2501.01957.

[14] CHEN J, HU Y, LI J, et al. Fireredchat: A pluggable, full-duplex voice interaction system with cascaded and semi-cascaded implementations[A]. 2025.

[15] WANG X, et al. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM[A]. 2024. arXiv: 2411.00774.

[16] CHEN Q, et al. Minmo: A multimodal large language model for seamless voice interaction [A]. 2025. arXiv: 2501.06282.

[17] WANG Z, et al. Neural-FSM: A full-duplex speech dialogue scheme based on large language model[C]//Proc. NeurIPS. 2024.

[18] XIE Z, et al. Mini-omni2: Towards open-source GPT-4o with vision, speech and duplex capabilities[A]. 2024. arXiv: 2410.11190.

[19] LU X, XU W, WANG H, et al. Duplexmamba: Enhancing real-time speech conversations with duplex and streaming capabilities[C]//CCF International Conference on Natural Language Processing and Chinese Computing. Springer, 2025: 62-74.

[20] MAI L, CARSON-BERNDSEN J. Real-time textless dialogue generation[A]. 2025.

[21] NGUYEN A D, et al. Generative spoken dialogue language modeling[J]. TACL, 2023.

[22] DÉFOSSEZ A, MAZARÉ L, ORSINI M, et al. Moshi: a speech-text foundation model for real-time dialogue[A]. 2024.

[23] VELURI A, et al. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents[A]. 2024. arXiv: 2409.15594.

[24] WANG Q, et al. NTPP: Generative speech language modeling for dual-channel spoken dialogue via next-token-pair prediction[A]. 2025. arXiv: 2506.00975.

[25] SHI Y, et al. Voila: A sophisticated, synchronous, and swift spoken language model[A]. 2025. arXiv: 2505.02707.

[26] YU D, et al. Salmonn-omni: A codec-free LLM for full-duplex speech understanding and generation[A]. 2025. arXiv: 2505.17060.

[27] HU K, et al. Salm-duplex: Efficient and direct duplex modeling for speech-to-speech language model[A]. 2025. arXiv: 2505.15670.

[28] WANG W, LI C, ZHANG L, et al. Covo-audio technical report[A]. 2026.

[29] TEAM T F, CHEN Q, CHENG L, et al. Fun-audio-chat technical report[A]. 2025.

[30] MA Z, et al. Language model can listen while speaking[A]. 2024. arXiv: 2408.02622.

[31] ZHANG J, et al. Omniflatten: A unified framework for spoken language model via progressive flattening[C]//Proc. ACL. 2025.

[32] LI Y, WU G, HOU H, et al. Uaf: A unified audio front-end llm for full-duplex speech interaction[A]. 2026.

[33] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[J]. Advances in neural information processing systems, 2020, 33: 12449-12460.

[34] HSU W N, BOLTE B, TSAI Y H H, et al. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units[J/OL]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460. DOI: 10.1109/TASLP.2021.3122291.

[35] CHEN S, WANG C, CHEN Z, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing[J/OL]. IEEE Journal of Selected Topics in Signal Processing, 2022, 16(6): 1505-1518. DOI: 10.1109/JSTSP.2022.3188113.

[36] RADFORD A, KIM J W, XU T, et al. Robust speech recognition via large-scale weak supervision[A]. 2022.

[37] YU H, CHEN Y, CAI M. Unit-based agent for semi-cascaded full-duplex dialogue systems [A]. 2026.

[38] CHEN Y, HE P, YU H, et al. Uniwhisper: Efficient continual multi-task training for robust universal audio representation[A]. 2026.

[39] LIN G T, LIAN J, LI T, et al. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities[A]. 2025.

[40] WANG C, YUE H, LI G, et al. Full-duplex interaction in spoken dialogue systems: A comprehensive study from the icassp 2026 humdial challenge[A]. 2026.

[41] STIVERS T, et al. Universals and cultural variation in turn-taking in conversation[J]. PNAS, 2009.

[42] LIN G T, et al. Full-duplex-bench v1.5: Evaluating overlap handling for full-duplex speech models[A]. 2025. arXiv: 2507.23159.

[43] LIN G T, KUAN S Y S, SHI J, et al. Full-duplex-bench-v2: A multi-turn evaluation framework for duplex dialogue systems with an automated examiner[A]. 2025.

[44] PENG Y, CHAO Y W, NG D, et al. Fd-bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems[A]. 2025.

[45] ZHANG H, CUI W, XU H, et al. Mtr-duplexbench: Towards a comprehensive evaluation of multi-round conversations for full-duplex speech language models[A]. 2025.

[46] ARORA S, LU Z, CHIU C C, et al. Talking turns: Benchmarking audio foundation models on turn-taking dynamics[A]. 2025.

[47] XU J, GUO Z, HU H, et al. Qwen3-omni technical report[A]. 2025.

[48] TEAM S. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier[J/OL]. GitHub repository, 2024. https://github.com/snakers4/silero-vad.

[49] WANG H, ZHENG S, CHEN Y, et al. Cam++: A fast and efficient network for speaker verification using context-aware masking[A].

[50] GAO Z, ZHANG S, MCLOUGHLIN I, et al. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition[A]. 2022.

[51] DENG W, ZHOU S, SHU J, et al. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system[A]. 2025.

[52] WANG X, LI Y, FU C, et al. Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen llm[A]. 2024.

[53] GENG X, SHAO Q, XUE H, et al. Osum-echat: Enhancing end-to-end empathetic spoken chatbot via understanding-driven spoken dialogue[A]. 2025.

[54] KIM C D, KIM B, LEE H, et al. AudioCaps: Generating captions for audios in the wild[C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019: 119-132. https://aclanthology.org/N19-1011/. DOI: 10.18653/v1/N19-1011.

[55] GEMMEKE J F, ELLIS D P W, FREEDMAN D, et al. Audio Set: An ontology and human-labeled dataset for audio events[C/OL]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2017: 776-780. DOI: 10.1109/ICASSP.2017.7952261.

[56] WU Y, CHEN K, ZHANG T, et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation[C/OL]//2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. https://arxiv.org/abs/2211.06687.

[57] MEI X, MENG C, LIU H, et al. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research[J/OL]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 3339-3354. DOI: 10.1109/TASLP.2024.3419446.

[58] BU H, DU J, NA X, et al. AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline[C]//2017 20th Conference of the Oriental Chapter of the

International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). 2017: 1-5.

[59] CHEN G, CHAI S, WANG G B, et al. GigaSpeech: An evolving, multi-domain ASR corpus with 10,000 hours of transcribed audio[C/OL]//Proceedings of INTERSPEECH 2021. 2021. https://www.isca-archive.org/interspeech_2021/chen21o_interspeech.html. DOI: 10.21437/Interspeech.2021-1965.

[60] GUAN S, LIU S, CHEN J, et al. Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays[C/OL]//2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Tokyo, Japan, 2021: 1116-1120. https://dblp.org/rec/conf/apsipa/GuanLCZLTYXCLWZ21.

[61] PANAYOTOV V, CHEN G, POVEY D, et al. LibriSpeech: An ASR corpus based on public domain audio books[C/OL]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2015: 5206-5210. DOI: 10.1109/ICASSP.2015.7178964.

[62] DROSSOS K, LIPPING S, VIRTANEN T. Clotho: An audio captioning dataset [C/OL]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020: 736-740. DOI: 10.1109/ICASSP40776.2020.9052990.

[63] CARTWRIGHT M, SEALS A, SALAMON J, et al. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations[J/OL]. Proceedings of the ACM on Human-Computer Interaction, 2017, 1(2). DOI: 10.1145/3134664.

[64] CARTWRIGHT M, CRAMER A, MENDEZ MENDEZ A E, et al. SONYC-UST-V2: An urban sound tagging dataset with spatiotemporal context[C/OL]//Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE). 2020. https://dcase.community/documents/workshop2020/proceedings/DCASE2020Workshop_Cartwright_68.pdf.

[65] HEITTOLA T, MESAROS A, VIRTANEN T. TAU urban acoustic scenes 2020 mobile,

development dataset[EB/OL]. 2020. https://zenodo.org/records/3670167. DOI: 10.5281/zenodo.3670167.

[66] URBAN-SED dataset[EB/OL]. 2018. https://zenodo.org/records/1324404. DOI: 10.5281/zenodo.1324404.

[67] CHEN H, XIE W, VEDALDI A, et al. VGGSound: A large-scale audio-visual dataset [C/OL]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020: 721-725. DOI: 10.1109/ICASSP40776.2020.9053174.

[68] XI Q, BITTNER R M, PAUWELS J, et al. GuitarSet: A dataset for guitar transcription[C]// Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR). 2018: 453-460.

[69] HAWTHORNE C, STASYUK A, ROBERTS A, et al. Enabling factorized piano music modeling and generation with the MAESTRO dataset[C/OL]//International Conference on Learning Representations (ICLR). 2019. https://openreview.net/forum?id=r1lYRjC9F7.

[70] BITTNER R M, SALAMON J, TIERNEY M, et al. MedleyDB: A multitrack dataset for annotation-intensive MIR research[C]//Proceedings of the International Society for Music Information Retrieval Conference (ISMIR). 2014.

[71] BOGDANOV D, WON M, TOVSTOGAN P, et al. The MTG-jamendo dataset for automatic music tagging[C/OL]//Machine Learning for Music Discovery Workshop (ML4MD), International Conference on Machine Learning (ICML). Long Beach, CA, USA, 2019. http://hdl.handle.net/10230/42015.

[72] AGOSTINELLI A, DENK T I, BORSOS Z, et al. MusicLM: Generating music from text [A/OL]. 2023. https://arxiv.org/abs/2301.11325. DOI: 10.48550/arXiv.2301.11325.

[73] THICKSTUN J, HARCHAOUI Z, KAKADE S M. Learning features of music from scratch[C/OL]//International Conference on Learning Representations (ICLR). 2017. https://openreview.net/forum?id=rkFBJv9gg.

[74] MANILOW E, WICHERN G, SEETHARAMAN P, et al. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity [C/OL]//2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). 2019: 45-49. DOI: 10.1109/WASPAA.2019.8937170.

[75] WECK B, MANCO I, et al. The song describer dataset: a corpus of audio captions for music-and-language evaluation[A/OL]. 2023. https://arxiv.org/abs/2311.10057. DOI: 10.48550/arXiv.2311.10057.

[76] MCKEE D, SALAMON J, SIVIC J, et al. Language-guided music recommendation for video via prompt analogies[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 14784-14793. https://openaccess.thecvf.com/content/CVPR2023/html/McKee_Language-Guided_Music_Recommendation_for_Video_via_Prompt_Analogies_CVPR_2023_paper.html.

[77] TURIAN J, SHIER J, et al. HEAR: Holistic evaluation of audio representations[C/OL]// Proceedings of Machine Learning Research: Vol. 176 Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track. PMLR, 2022: 120-136. https://proceedings.mlr.press/v176/turian22a.html.

[78] ZHANG J, DINKEL H, NIU Y, et al. X-ARES: A comprehensive framework for assessing audio encoder performance[A]. 2025.

[79] BAEVSKI A, ZHOU Y, MOHAMED A, et al. wav2vec 2.0: A framework for self-supervised learning of speech representations[C/OL]//Advances in Neural Information Processing Systems: Vol. 33. 2020. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

[80] CHEN S, WU Y, WANG C, et al. BEATs: Audio pre-training with acoustic tokenizers [C/OL]//Proceedings of Machine Learning Research: Vol. 202 Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023: 5178-5193. https://proceedings.mlr.press/v202/chen23ag.html.

[81] ELIZALDE B, DESHMUKH S, AL ISMAIL M, et al. CLAP: Learning audio concepts from natural language supervision[C/OL]//ICASSP 2023. 2023: 1-5. DOI: 10.1109/ICASSP49357.2023.10095889.

[82] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C/OL]//Proceedings of Machine Learning Research: Vol. 139 Proceedings of the 38th International Conference on Machine Learning. PMLR, 2021: 8748-8763. https://proceedings.mlr.press/v139/radford21a.html.

[83] WU Y, CHEN K, ZHANG T, et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation[C/OL]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5. DOI: 10.1109/ICASSP49357.2023.10095969.

致 谢

我未曾不每日感谢命运所带来的一切。行至今天这般，父母无条件的支持，蔡敏捷老师提供的科研机会，给到学业生活帮助的师兄朋友同学，提供实习推荐的朋友，在长沙一同创造美好回忆的朋友，工作中的朋友，搭建这个世界的一切相遇不曾相遇过的人，我未曾不怀有深刻的感激之心。是你们构建了我的生命的织机。

浮世景色百千年依旧。感谢这个世界提供的无法预测的命运之舞台。

附录 A　签名