arXiv:2603.08216v1 [eess.AS] 9 Mar 2026

DualTurn：从双通道生成式语音预训练中学习轮替

Shangeth Rajaa[1]

[1] Anyreach AI

shangeth@anyreach.ai

摘要

语音到语音模型能够自然地处理轮替，但对工具调用或复杂推理的支持有限；而生产环境中的 ASR–LLM–TTS语音流水线虽然具备这些能力，却依赖静音超时，导致轮替不自然。我们提出 DualTurn，通过对双通道对话音频进行生成式预训练，缩小这一差距。该模型以自回归方式生成两位说话人接下来的音频，在无需任何标签的情况下隐式学习对话动态，随后通过微调来预测可解释的轮替信号，并将其直接映射到智能体动作。DualTurn 持续监控两个通道，预判轮替边界并生成五种智能体动作。在标准基准测试中，DualTurn（0.5B）在智能体动作预测方面优于 VAP（加权 F1 为 0.633 vs. 0.389），在词级轮次预测方面优于一个 3.1B 音频文本模型（曲线下面积为 0.930 vs. 0.880），同时能更早预判轮替边界并减少打断。

索引词：口语对话系统、轮替、对话式 AI、自监督学习、音频语言模型

1. 引言

基于大语言模型（LLM）构建的生产级语音流水线利用工具调用和推理能力，但仍依赖静音超时来控制轮替，这可能导致响应延迟和打断。近期的语音端点检测模型也使用声学特征[1] 和多模态特征 [2] 来解决这一问题；但由于它们依靠语音活动检测（VAD）来识别停顿并触发，因此无法预测语音何时结束，只能被动响应。语音到语音（S2S）模型的生成式训练 [3, 4, 5, 6] 要求根据说话人声道预测下一音频帧，这使模型能够预测任一说话人接下来会说什么，从而隐式地学习轮替动态。然而，S2S 模型缺乏文本大语言模型的推理能力和指令遵循能力，其轮替能力也无法迁移到模块化 ASR-LLM-TTS 流水线中。

尽管轮替发言研究 [7, 8] 已取得显著进展，但目前还没有模型能够涵盖真实生产流程所需的全部现象。基于文本的模型 [9] 根据转录文本预测轮次是否结束，却忽略了轮替发言所必需的韵律信息 [10, 11]。音频分类器 [12, 13]虽能纳入韵律信息，但仅限于单通道二分类；由于缺少另一位说话人的上下文，它们只能检测轮次结束，无法判断轮次动作。反馈语 [14], 、打断、轮次中停顿和语音重叠 [15] 都是会话轮替发言中自然出现的现象。Wang 等人 [16] 使用一个拥有 31 亿参数的大语言模型，利用音频和转录文本输入对轮次转换、反馈语和保持进行分类，不过这类模型对于低延迟部署而言仍然过于庞大——延迟部署而言仍然过于庞大。然而，仅检测语音何时结束还不够。模型还必须预测说话人的行为。语音活动投射（VAP） [17] 最接近这一目标，它是一种自监督双通道模型，仅根据音频持续预测未来的语音活动。但 VAP 将所有现象都简化为二元语音活动概率，无法区分反馈语和轮次结束；其拥有 580 万参数的 CPC 编码器缺乏语义建模能力，而且还需要额外微调，才能使反馈语预测 [18]的表现超过随机水平。

我们提出 DualTurn，这是一种轮替组件，可用于基于双通道音频 S2S 生成式预训练的 ASR-LLM-TTS 流水线。由于 DualTurn 同时对两位说话人进行建模，因此能够获取单通道模型无法捕捉的会话上下文，例如语音重叠、打断和反馈语。它通过学习预测任一说话人接下来会说什么，无需任何标签即可掌握轮替和会话概念。模型使用冻结的 Mimi 神经编解码器对音频进行编码； [3], 0.5B 大语言模型骨干网络接收来自 Mimi 编解码器的音频嵌入，并通过多层感知机头为每个说话人通道预测六种分类信号。这些信号既可以通过启发式方法组合成智能体动作，也可以通过逻辑回归探针进行组合。该模型可以持续监控两位说话人的通道，并在单个 CPU 上运行。

在 Switchboard 和 otoSpeech 上的所有评估设置中，DualTurn 的表现均优于 VAP。DualTurn 在词级预测方面优于 3.1B 参数的音频-文本融合模型， [16] 并能在语音结束前预判轮替边界，同时减少误判打断。我们的主要贡献在于展示：据我们所知，首次将双通道会话音频的 S2S 生成式语音预训练用作模块化流水线中显式轮替预测的表征学习阶段，能够产生非常适合轮替信号预测的表征。我们从这些表征中提取每通道六种信号（轮次结束、轮次中停顿、语音起始、反馈语、语音活动和未来活动），并将其组合成五种智能体动作。模型持续运行，而不必等待语音活动检测发出停顿信号。

2. 方法

2.1. 架构

DualTurn 使用 Mimi 神经编解码器 [3] 对双通道音频流中的两位说话人进行编码，将每个通道的 24 kHz 波形转换为 8 条残差向量量化码本序列，速率为 12.5帧·秒$^{-1}$。我们使用连续编码器嵌入（每通道 512 维），而非离散码本索引。每个通道的特征都经过通道专属多层感知机；

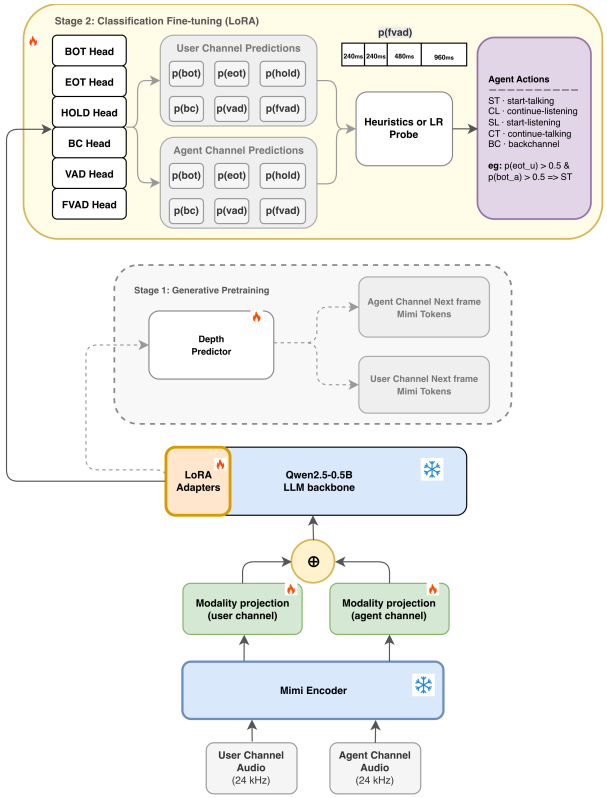

图1： *DualTurn* 架构。阶段一对骨干网络和深度预测器进行预训练（预训练后丢弃），以预测下一个语音词元；第二阶段微调十二个分类头，以预测轮流发言信号。

两条表征被拼接后送入 Qwen2.5-0.5B [19]，，并按下文所述分两个阶段进行训练。十二个轻量级分类头（每个通道六个）连接到骨干网络的最终隐藏状态。稀疏信号（EOT、保持、BOT、反向通道）由带有 GELU 激活和丢弃法的两层多层感知机头预测，而稠密信号（语音活动检测、帧级语音活动检测）仅需线性投影。帧级语音活动检测头各自输出四个时间范围值，使 12 个头的标量输出总数达到 18（两个帧级语音活动检测头各产生四个时间范围值；其余头各产生一个）。（图 1）。推理期间，模型处理来自两个说话人声道的流式音频，并以 240 ms 步长（每次推理 3 个音频帧）结合 KV缓存进行预测。基础未量化模型的延迟在 CPU 上约为 78ms，在 A100 GPU 上约为 27ms。

2.2. 阶段一：生成式语音预训练

骨干网络经过预训练，可同时自回归地预测两位说话人的下一个音频词元。轻量级深度预测器（∼10.6M 个参数）根据骨干网络的输出，生成每个通道下一帧的 RVQ 码本词元。阶段一完成后，深度预测器将被丢弃，仅保留表示骨干网络用于第二阶段训练。

生成训练损失迫使模型学习并理解语义、韵律信息和交互模式，这些信息随后可用于第二阶段微调中的轮替预测。

表1： 自监督信号定义（每个通道） *l)*

信号	定义
EOT	语音结束点；另一位说话人开始发言在 4 秒内
HOLD	所有其他语音结束点（非 EOT）
BOT	另一位说话人之后的语音起始（≥1 秒）
BC	说话人独立话语 ≤1 秒，≥1秒静音之前/之后之前/之后
VAD	逐帧二值语音活动
FVAD	在 [0–240 毫秒], [240– 的平均语音活动480 毫秒], [480–960 毫秒], [960毫秒–2 秒] 之前

表 2： 智能体动作和真实标注定义。

动作	定义
ST 开始说话	用户语音结束；智能体在 4 秒内开始说话
CL 继续聆听	用户语音结束；用户在 2 秒内继续说话
SL 开始聆听	语音重叠起始；传入语音 >1秒
CT 继续说话	语音重叠起始；传入语音 <1秒
BC 应答语	用户说话期间智能体发声 <1秒语音

2.3. 第二阶段：轮替信号预测

第二阶段中，预训练骨干网络会预测每个通道的六种轮流发言信号。这些信号的标签完全通过自监督语音活动对齐得出，无需任何人工标注（表1）。

我们采用 4 秒的前瞻时长（而 VAP 为 1 秒），以免丢弃 Switchboard 中 12% 的话轮交接案例；这些案例的停顿超过 1 秒。EOT 和 HOLD 在设计上互为补充：在每个语音结束点，恰好有一个信号触发，因此模型无需任何标注即可区分轮次结束与轮次中停顿。对于稀疏信号，我们采用焦点损失 [20]；对于语音活动检测和帧级语音活动检测，则采用二元交叉熵。标签通过非对称高斯函数进行时间平滑（每个事件前 σ=3 帧、后 σ=1 帧），从而训练模型提前最多 240 毫秒预测每个信号。

2.4. 智能体动作推断

预测信号（每个通道六个）被映射为五种智能体动作（表 2）。ST 和 CL 分别对应 VAP 的切换和保持决策，而 SL、CT 和 BC 则捕捉现有模型未处理的语音重叠和反向通道事件。作为零参数基线，领域知识启发式规则直接对各个信号设定阈值（例如 $\text{EOT}_u > 0.5 \wedge \text{BOT}_a > 0.5 \Rightarrow$ ST）。为根据预测信号做出更丰富的类别决策，我们还使用在留出验证数据上拟合的多项逻辑回归（LR），将信号概率映射为智能体动作。学习得到的 LR 系数具有可解释性。例如，开始说话动作更关注 EOT_u 和 VAD_a（用户轮次已结束，智能体正在说话），同时对 VAD_u 和 BOT_u 取反（用户静默，且不会重新开始说话）。LR 将语音活动、未来活动和事件信号（EOT、BOT、HOLD、BC）组合为单一的智能体动作。

3. 实验

3.1. 设置

我们使用了约 453 小时的双通道会话音频来预训练模型。该数据集由 oto-Speech [21]（1,125 段英语全双工在线会话，共 289 小时，采样率为 24kHz）和 Switchboard [22]（220 小时、采样率为 8kHz 的电话语音，其中 138 个会话的测试集划分留作评估）组成。在阶段一，我们以 lr=3×10^{-4} 训练 LoRA 适配器 [23]，随后以 lr=10^{-5} 对大语言模型参数进行全量微调，并在单张 A100 40GB GPU 上采用批量大小为 32 的设置进行早停。在第二阶段，我们仅以 lr=1×10^{-4} 微调预测头和 LoRA 适配器，最多训练 10 轮，并采用早停且舍弃生成损失。两个阶段中，Mimi 编解码器均保持冻结。

Switchboard 数据集主要用于评估，并在先前研究中使用的同一 138 场次测试集划分上进行评估 [17, 16]。otoSpeech 则在一个不重叠的 113 场次测试集划分上进行补充评估。外部基线包括 VAP [17]（580 万个参数、CPC 编码器、双通道音频）以及已发表的最佳音频+文本系统 [16]（最多 31 亿个参数、RedPajama +HuBERT）。第 3.4 节定义的消融变体 A–G 列于表 3 中。

使用启发式方法或逻辑回归探针，根据 12 个信号头预测的信号推断智能体动作（第 2.4 节）。该探针只能对神经网络模型已捕获的信号或概率进行线性聚合。评估时采用原论文中的定义：在 VAP 帧级协议 [17] 以及 Wang 等人 [16], 的设置下，均使用相同的 Switchboard 测试集进行评估。表 3 中的所有模型均在相同的 4 秒动作定义下进行评估。表 4 进一步在 VAP 原始的 1 秒条件下评估 DualTurn，以确认性能提升源自学到的表征，而非标签定义。

3.2. 主要结果

Wang 等人 [16] 提供了与词级基准测试最直接的外部比较。学习到的表征足够强，即使不使用聚合器，单独的 EOT 信号（平均曲线下面积为 0.914）也优于 Wang 等人的 3.1B 音频+文本模型（平均曲线下面积为 0.880），且无需与其他信号结合。结合多个学习到的信号后，性能进一步提升至平均曲线下面积 0.930，而逻辑回归探针取得了最佳性能，平均曲线下面积达到 0.963（表 5）。

VAP（原生）将 VAP 自身的 p_{now} 和 p_{future} 信号直接用于智能体动作决策，而 VAP（LR-6）则在验证集上对 VAP 的六个投影输出拟合逻辑回归。DualTurn 在两个评估数据集的全部五种智能体动作类别上均优于 VAP（Switchboard 上加权 F1 为 0.633 对 0.389，otoSpeech 上加权 F1 为 0.707 对 0.461），见表 3。应答检测的差距最大，VAP 的 BC F1 = 0.000，这是因为它没有专门的 BC 信号 [18]。即使使用逻辑回归探针对其预测的投影概率进行分析，VAP 也无法区分应答语和继续交谈。与 VAP 一样，DualTurn 完全使用自监督标签训练，在预测“智能体现在应当发出应答语”时，BC F1 = 0.349（随机水平 F1 ≈ 0.080）。这证实了阶段一生成式预训练的优势。表 3 显示，经过阶段一预训练和未经预训练的模型在 BC F1 上呈现明显的二元分化；第 3.3 节将对此进行分析。

DualTurn 在全部四项任务中均优于 VAP（表 4），其中话轮转移/保持的提升最大（+0.142 加权 F1），短/长、话轮转移预测和 BC 预测也展现出持续的改进。尽管 DualTurn 是按照我们更宽泛的 4 秒定义训练的，但在 VAP 自身的评估方案下，仍取得了这些改进。

表 3：*Switchboard*（*138* 场会话）和 *otoSpeech*（*113* 场会话）上的智能体动作预测。*wF1* = 加权 *F1* 分数（*5* 类）。*Ant@−240* = 结束点前 *240* 毫秒的 话轮转移/保持*AUC*（随机水平为 *= 0.500*）。变体 *A–G* 的定义见第 *3.4* 节。

模型	wF1	BC F1	Ant@−240
— *Switchboard* —			
VAP（原生）	.276	—	.785
VAP（LR-6）	.389	.000	.780
A – LoRA（主要）	**.633**	**.349**	.874
B – LoRA + CB	.613	.077	.876
C – 无预训练	.604	.079	.863
D – 长短期记忆网络	.602	.077	.868
E – 全量微调	.626	.337	**.879**
F – 离散	.602	.072	.867
G – 文本感知型 PT	.605	.085	.873
— *otoSpeech* —			
VAP（LR-6）	.461	.000	.719
E – 全量微调	**.709**	.498	**.860**
A – LoRA（主）	.707	**.512**	.848

表 4：在 *Switchboard* 上使用帧级 *VAP* 协议，并依据 *VAP* 原始的 *1*秒双向事件条件 *[17]*进行评估。所有指标均为加权 *F1*。*S/H* = 话轮转移/保持，*S/L* = 短/长，*S-P* =话轮转移预测，*BC-P* = *BC* 预测。*VAP* 论文采用了 *11* 折交叉验证（*S/H = 0.899*）；我们的固定划分重新评估结果为 *0.843*。

模型	S/H	S/L	S-P	BC-P
VAP（我们的评估）	.843	.916	.720	.838
A – LoRA	**.985**	.979	.764	.864
E – 全量微调	.982	**.985**	**.771**	**.869**

DualTurn 能比 VAP 提前约 220 毫秒检测到轮替边界，相对于轮次结束的中位反应时间分别为 −360 毫秒和 −140 毫秒（图2）。它也较少误触发：ST-for-CL 混淆率从 27.4% 降至 22.4%，ST F1 分数从 VAP 的 0.808 提升至 0.829，打断减少了 5 个百分点。

3.3 分析

语音预训练开启反馈信号动作预测。表 3 显示，反馈信号表现存在明显分化：所有模型都根据是否经过阶段一预训练分为两类。其他所有模型的 BC F1 分数都不超过 0.085，而模型A和E是仅有的两个经过阶段一预训练、且具有纯净阶段2分类目标的模型，其 BC F1 分数达到 0.337–0.349。BC 的提升幅度显著，而非渐进式增长：召回率从 0.045 提升至 0.458，F1 分数相对提升了 +340%。不过，由于反馈信号在评估数据的所有事件中占比不足 8%，精确率仅为 0.282（高于 0.080 的随机水平），这表明该信号确实存在，但效果有限。预训练使得即使在数据如此稀疏的情况下，也能预测智能体反馈信号动作，但尚未彻底解决这一问题。部署时，BC 信号不应作为唯一的触发条件；相反，它应采用更高的阈值，或作为软输入来引导

表 5： *Switchboard* 上的词级轮次预测。*C* =继续交谈，*B* = 应答语，*T* = 轮次转换。基线来自 *[16]*。*A–EOT*：单一 *EOT* 信号，*0* 个参数。*A–Heur*：多信号启发式方法，*0* 个参数。*A–LR*：对所有信号头进行逻辑回归探针。

模型	曲线下面积(C)	曲线下面积(B)	曲线下面积(T)	平均	EER
GPT-2	.851	.774	.862	.829	24.5
RP+HuBERT+历史	.903	.818	.920	.880	19.3
A – 仅 EOT	.918	.904	.919	.914	15.2
A – 启发式方法	.940	.925	.924	.930	13.2
A – 逻辑回归探针	**.961**	**.979**	**.950**	**.963**	**9.7**

决策策略。

仅靠架构无法复现这一结果。8M 长短期记忆网络模型（D）和 0.5B 大语言模型（C）的表现几乎相同：加权 F1 分数分别为 0.602 和 0.604，BC F1 分数分别为 0.077 和 0.079。总 BC F1 分数提升中，超过 99% 来自预训练（A 与 C 的比较），而非架构（D 与 A 的比较）。大语言模型的贡献仅约为 +0.002 BC F1。没有预训练时，大语言模型相较于长短期记忆网络的容量优势微乎其微；但要让预训练发挥作用，又需要足够的容量。骨干网络并非轮替知识的来源，而是承载预训练所注入知识的必要载体。

语义和韵律驱动轮替。 第二阶段的轮替预测信号分为两种难度类别。语音活动检测和帧级语音活动检测是密集的逐帧信号，所有变体（包括长短期记忆网络）都能轻松学会。BOT、反向通道以及在较小程度上的 EOT，则是稀疏的交互事件，需要根据有限的训练信号识别双说话人的动态。这一差距通过预训练得以弥合：BOT 提升了 +188%，反向通道提升了 +41%，而语音活动检测和帧级语音活动检测的提升则微乎其微。我们对训练后的离散模型 F 进行了码本消融，方法是将单个码本置零或单独隔离。结果显示，各码本对话轮转移/保持AUC的贡献排序一致。话轮结束判别信号的 56% 来自码本 0（语义），码本 1 贡献 26%，其余 18% 则由码本 2–7 共同贡献。语义和韵律信息的贡献超过 80%，精细声学特征的贡献微乎其微。这也与先前的研究结果一致 [10]。

预训练形成多尺度注意力层级。我们通过探查大语言模型各个注意力头/层所关注的平均时间距离，对训练后的 DualTurn 模型进行了注意力分析。结果显示，大语言模型有 6 层关注短距离（<1 秒），可能负责处理帧级声学细节；另有 3 层关注长距离（L7 = 14.9 秒、L9 = 12.6 秒、L11 = 14.8 秒），以在时间上维持对话语境

这是 LSTM 无法实现的规模。这解释了为什么 C≈D 的表现更好：没有阶段一预训练时，模型 C 只有 3 层关注短距离范围，长距离注意力则较为分散。经过预训练的 Transformer 还会在没有显式监督的情况下形成特定于动作的行为。对于 CL 类，用户通道受到关注的程度高出 3.77×；对于 BC 类，则高出 2.21×。各动作类别的时间范围也有所不同：BC 为 3.41 秒，CT 为 4.91 秒。

3.4. 消融研究

表 3 包含七种变体：**A** （主配置；LoRA、预训练、连续 Mimi、Transformer），**B** （在第二阶段保留辅助生成损失），**C**（无阶段一预训练），**D** （800 万

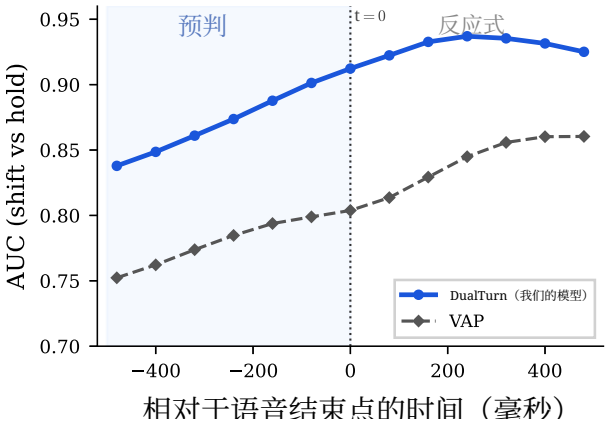

图2：相对于语音结束点各时间偏移处的切换与保持 *AUC*（*Switchboard* 测试集）。

参数的长短期记忆网络骨干网络），**E** （对全部 5 亿个参数进行完整微调），**F** （离散码本索引），以及**G** （在阶段一加入文本生成式 ASR 目标）。

预训练和骨干网络的影响已在第 3.3 节讨论（A 与 C、A 与 D 与 C）。尽管参数更少，LoRA（A）在 BC F1（0.349 对 0.337）和加权 F1（0.633 对 0.626）上都略胜于全量微调（E）。我们比较了在阶段一预训练条件下，使用离散码本索引（F）与连续 Mimi 表征（A）作为输入的效果，发现连续输入表征在所有指标上均优于离散码本输入（加权 F1 0 1 0 633 0 602，349 0 072 . 对 . BC F . 对 . ），因为将连续表征量化为离散索引会破坏韵律细微差异，而且离散模型必须从头学习码本嵌入。

我们还设计了两项负向消融实验。将阶段一的生成损失作为辅助损失，会使 BC F1 从 0.349 降至 0.077（A 与 B 对比），加权 F1 从 0.633 降至 0.613，因为相互竞争的梯度抑制了对稀疏任务（EOT、BOT、保持、反向通道）的学习。为利用大语言模型预训练获得的文本生成能力，并促使参数学习语音-文本模态适配，我们在阶段一加入了文本输出目标（自动语音识别任务），并观察到第二阶段的性能有所下降（A 与 G 对比：加权 F1 为 0.633 vs. 0.605，BC F1 为 0.349 vs. 0.085）。这表明，与文本对齐的预训练相比，纯音频模态训练能为轮流发言任务学习到更好的表征。

4. 结论

在会话音频上对双通道生成式语音模型进行预训练，并针对显式轮替发言信号进行微调，可以缩小基于静音的端点检测方法与语音到语音级轮流发言动态之间的差距。在无需人工标注的情况下，DualTurn 的表现优于 VAP 和 3.1B 音频-文本融合模型等其他方法，展现出更强的预判能力，其预测轮次结束的时间比 VAP 早 220 毫秒。预训练增强的多尺度注意力层次结构使模型能够预测智能体反馈语。将规模扩展到更大规模的多语言/多方语料库，并优化生成式预训练，是自然的后续步骤，因为在我们的实验中，训练和评估仅限于 453 小时的英语双人对话语音。主要结论是，对于轮替发言而言，生成式预训练是教师，大语言模型骨干网络是承载知识的载体，而不是知识来源。

5. 参考文献

[1] S. Udupa、K. Kumar、S. Majumdar、J. Balam 和 B. Ginsburg，《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》，arXiv 预印本 arXiv:2506.07081，2025。[2] 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：融合声学和语言模态，实现全双工口语对话系统中的稳健轮替发言》，2025。 [在线]。可访问：https://arxiv.org/abs/2509.23938[3] A. D´efossez、L. Mazar´e、M. Orsini、A. Royer、P. P´erez、H. J´egou、E. Grave 和 N. Zeghidour，《Moshi：面向实时对话的语音文本基础模型》，2024。[4] 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》，《计算语言学协会学报》，第 11 卷，第 250–266 页，2023。[5] R. Roy、J. Raiman、S. gil Lee、T.-D. Ene、R. Kirby、S. Kim、J. Kim 和 B. Catanzaro，《PersonaPlex：全双工会话语音模型中的语音和角色控制》，2026。[6] Q. Wang、Z. Meng、W. Cui、Y. Zhang、P. Wu、B. Wu、I. King、L. Chen 和 P. Zhao，《Ntpp：通过下一词元对预测实现双通道口语对话的生成式语音语言建模》，2025。 [在线]。可访问：https: //arxiv.org/abs/2506.00975[7] H. Sacks、E. A. Schegloff 和 G. Jefferson，《会话轮替发言组织的最简系统》，Language，第 50 卷，第 4 期，第 696–735 页，1974。[8] G. Skantze，《会话系统与人机交互中的轮替发言：综述》，Computer Speech & Language，第 67 卷，第 101178 页，2021。[9] E. Ekstedt 和 G. Skantze，《TurnGPT：用于预测口语对话中轮替发言的基于 Transformer 的语言模型》，载于《计算语言学协会论文集：EMNLP 2020》，2020 年，第 2981–2990 页。[10] A. Gravano 和 J. Hirschberg，《面向任务的对话中的轮替发言线索》，Computer Speech & Language，第 25 卷，第 3 期，第 601–634 页，2011。[11] N. Ward 和 W. Tsukahara，《提示英语和日语中应答语回应的韵律特征》，Journal of Pragmatics，第 32 卷，第 8 期，第 1177–1207 页，2000。[12] M. Roddy、G. Skantze 和 N. Harte，《使用长短期记忆网络研究用于连续轮替发言预测的语音特征》，载于 Interspeech 2018 论文集，2018 年，第 586–590 页。[13] A. Raju、A. Khare、D. He、I. Sklyar、L. Chen、S. Alptekin、V. A. Trinh、Z. Zhang、C. Vaz、V. Ravichandran、R. Maas 和 A. Ras-trow，《用于语音识别的双遍端点检测》，载于 IEEE 自动语音识别与理解研讨会（ASRU），2023 年，第 1–8 页。[14] R. R¨ude、M. M¨uller、S. St¨uker 和 A. Waibel，《利用词嵌入增强反馈语预测》，载于 Interspeech 2017 论文集，2017 年，第 879–883 页。[15] M. Heldner 和 J. Edlund，《会话中的停顿、间隙和重叠》，Journal of Phonetics，第 38 卷，第 4 期，第 555–568 页，2010。[16] J. Wang、L. Chen、A. Khare、A. Raju、P. Dheram、D. He、M. Wu、A. Stolcke 和 V. Ravichandran，《利用声学与大语言模型融合进行轮替发言和反馈语预测》，载于 IEEE 国际声学、语音与信号处理会议（ICASSP）论文集，2024。[17] E. Ekstedt 和 G. Skantze，《语音活动投影：轮替发言事件的自监督学习》，载于 Interspeech 2022 论文集，2022 年，第 5190–5194 页。[18] K. Inoue、D. Lala、G. Skantze 和 T. Kawahara，《Yeah、un、oh：通过微调语音活动投影进行连续实时反馈语预测》，载于 NAACL 2025 论文集，2025。

[19] Qwen 团队，“Qwen2.5 技术报告”，*arXiv* 预印本 *arXiv:2412.15115*，2025。[20] T.-Y. Lin、P. Goyal、R. Girshick、K. He 和 P. Doll´ar，“用于密集目标检测的焦点损失”，载于*IEEE* 国际计算机视觉会议（*ICCV*）论文集，2017。[21] otoearth，“otoSpeech-full-duplex-280h：全双工对话语音数据集”，https://huggingface.co/datasets/otoearth/otoSpeech-full-duplex-280h，2025，许可协议：知识共享署名 4.0 国际许可协议。[22] J. J. Godfrey、E. C. Holliman 和 J. McDaniel，“Switchboard：用于研究与开发的电话语音语料库”，载于*IEEE* 国际声学、语音与信号处理会议（*ICASSP*）论文集，1992，第 517–520 页。[23] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang 和 W. Chen，“LoRA：大语言模型的低秩适配”，载于国际学习表征会议（*ICLR*）论文集，2022。 [在线]。获取地址：https://arxiv.org/abs/2106.09685