从轮流发言到同步对话：全双工口语模型综述

Yuxuan Chen[1*] Haoyuan Yu[2]

[1] Jilin University, Changchun, China [2] Hunan University, Changsha, China
yxchen5522@jlu.edu.cn y15352176976@hnu.edu.cn

ABSTRACT

真正全双工（TFD）语音通信——实现同时倾听和说话，自然的轮流发言、重叠言语和打断——代表了朝向类人类 AI 交互的关键里程碑。本综述全面回顾了 LLM 时代中的全双工口语语言模型（FD-SLMs）。我们建立了一个分类法，区分了工程同步（模块化架构）与学习同步（端到端架构），并将零散的评估方法统一为一个框架，涵盖时间动态、行为仲裁、语义连贯性和声学性能。通过对主流 FD-SLMs 的比较分析，我们识别出基本挑战——同步数据稀缺、架构分歧和评估差距——提供了推进人机通信的发展路线图。

对于代码和更多细节，请参见 GitHub [1]。

Index Terms— 真正的全双工，全双工口语模型，认知并行性，同步

1. 介绍

当代 SLMs 从根本上缺乏进行自然对话所必需的同时听和说的能力。虽然 LLMs 在语言理解上实现了革命性突破 [1,2]，但它们的口语对话实现仍然受限于顺序的听-思考-说循环。当前系统通过时分复用仅实现伪全双工（PFD）行为，未能匹配以自然轮流说话行为为特征的人类对话动态 [3,4]，如图 1所示。

FD-SLMs 将这一范式从顺序认知架构转变为并行认知架构。与 PFD 系统在倾听和说话之间交替不同，FD-SLMs 能够在同一处理周期内同时进行编码和生成，支持包括打断、反馈渠道以及通过双向信息流实现的自适应轮流发言在内的自然对话事件。

[1]https://github.com/elpsykongloo/FD-SLMs

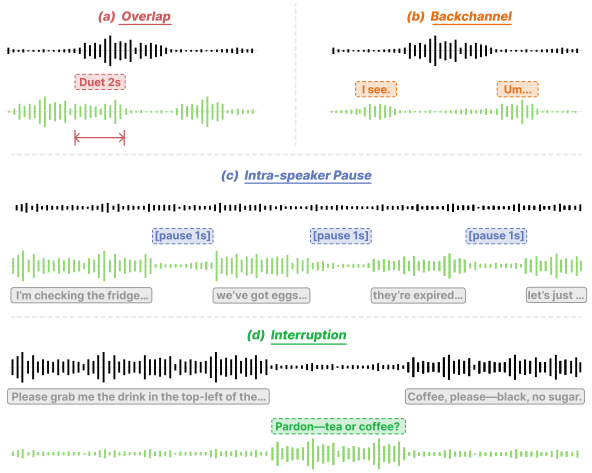

Fig. 1. 自然对话包含轮流事件：（a）重叠，（b）回渠，（c）停顿，和（d）打断。

早期系统展示了增量处理 [5] 和有限状态控制 [6]，实现了响应性而没有语义灵活性。LLM 集成带来了工程化同步 [7–9] 和端到端架构。在 dGSLM 的新兴轮流发现 [10] 之后，最近的进步包括层次多流处理 [11]、下一令牌对预测（NTPP）[12] 以及连续离散对齐 [13]。

尽管取得了这些进展，现有的调查 [14,15] 仍将全双工视为实现细节而非基本要求，缺乏系统性的 FD-SLM 设计分析。评估仍然支离破碎 [16–18]。

本文作出了以下主要贡献：

1. **正式双重特征描述：** 数学定义严格区分了半双工、伪全双工和真正的全双工系统，揭示了认知并行计算的需求。
2. **建筑分类学：** 系统分类揭示了在同步策略、状态管

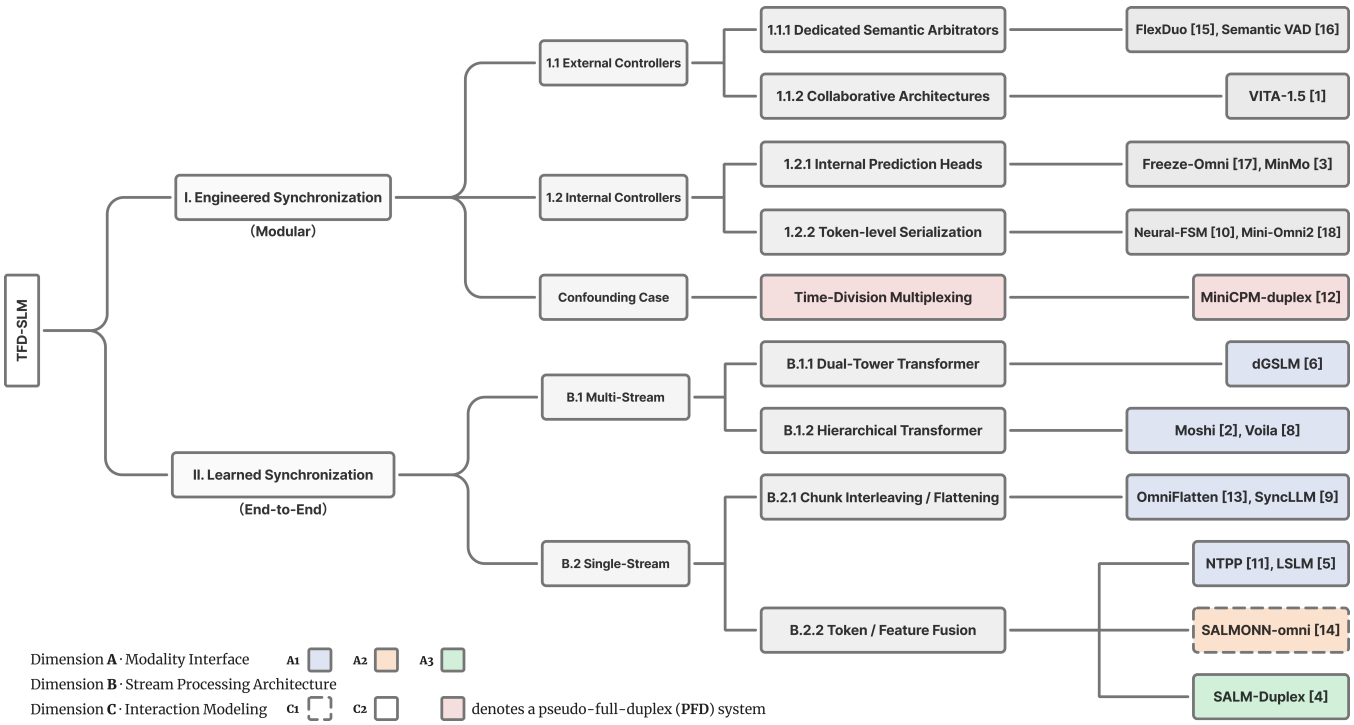

Fig. 2. 开源 FD-SLMs 的架构分类法。

理和训练范式轴上的设计空间，识别了权衡和未探索的机会。

3. **系统评价和分析：**我们比较了七种 FD-SLM 的流架构，识别出关键的数据瓶颈，并建立了一个四支柱基准分类法，揭示了当前系统所面临的根本性的延迟与质量之间的权衡。

2. 形式化

FD-SLMs 实现认知并行性，通过同时编码输入和解码输出，实现实时输出适应。让代理 $\mathcal{A}$ 与环境 $\mathcal{E}$ 进行交互。由于直接建模连续音频 $X(t)$ 是不可行的，离散化器 $\mathcal{T}$ 生成对齐序列 $S^{\mathcal{E}} = (e_1, \ldots, e_T)$ 和 $S^{\mathcal{A}} = (a_1, \ldots, a_T)$，将 e_t 与 a_t 对齐以实现同步交互。

2.1. 联合概率视角

交互建模范式建模联合分布 $P(S^{\mathcal{E}}, S^{\mathcal{A}})$：

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

这构成了 NTPP [12] 的基础，在仅解码器的变压器中同时预测 (e_t, a_t) 对：

$$\mathcal{L}_{\mathrm{NTPP}}(\theta) = \mathbb{E}_{(S^{\mathcal{E}}, S^{\mathcal{A}})}\Big[\textstyle\sum_{t=1}^{T} \log P\big(e_t, a_t \mid S^{\mathcal{E}}_{<t}, S^{\mathcal{A}}_{<t}; \theta\big)\Big] \tag{2}$$

早期的方法 [10] 通过交叉注意力中的条件独立性进行近似，优化的是加和的条件对数似然而非真实的联合似然。

2.2. 条件概率视角

对于交互式智能体，目标变成建模 $P(S^{\mathcal{A}} \mid S^{\mathcal{E}})$：

$$a_t \sim P\big(a_t \mid S^{\mathcal{E}}_{\le t}, S^{\mathcal{A}}_{<t}; \theta\big)\,. \tag{3}$$

并发。在摄取 $e_{t+1}, \ldots$ 的同时计算 a_t 需要并行编码-解码 [11, 19]。

实时约束。$\mathrm{Time}(\mathrm{Compute}(a_t)) < 200\,\mathrm{ms}$ [3].

自我调节。对 $S^{\mathcal{A}}_{<t}$ 的依赖性确保了相干性并能够实现回声消除 [13]。

训练目标：

$$\mathcal{L}_{\mathrm{Cond}}(\theta) = \mathbb{E}_{(S^{\mathcal{E}}, S^{\mathcal{A}})}\Big[\textstyle\sum_{t=1}^{T} \log P\big(a_t \mid S^{\mathcal{E}}_{\le t}, S^{\mathcal{A}}_{<t}; \theta\big)\Big] \tag{4}$$

使用同步数据进行训练，无论采用哪种目标函数，都能使轮流动态在没有监督的情况下出现。

Table 1. 开源 FD-SLM 中建筑组件的比较分析。

模型	输入感知	核心处理	输出合成
dGSLM	HuBERT + k-means clustering	Two-tower Transformer with cross-attention	HiFi-GAN unit vocoder
如果	Mimi neural codec (RVQ)	RQ-Transformer joint autoregression	Mimi decoder
同步大语言模型	HuBERT features	Interleaved and predictive synchronization	HiFi-GAN vocoder
萨尔蒙全能型	Mamba streaming encoder	Dynamic control tokens for stream management	CosyVoice2 with fixed-length generation
最小模型	SenseVoice-Large + projector	Full-Duplex Predictor (FDP) head	CosyVoice2 chunk-aware flow-matching
柔性双组件系统	Qwen2-Audio encoder	Finite-state machine control	External TTS delegation
VITA 1.5	Conv + Transformer encoder	Dual LLM instances with shared KV cache	TiCodec decoder

2.3. 分层和预测机制

分层生成。像 Moshi 这样的系统 [11] 利用文本表示 $T^{\mathcal{A}}$：

$$P(S^{\mathcal{A}} \mid S^{\mathcal{E}}) = \int P\big(S^{\mathcal{A}} \mid T^{\mathcal{A}}, S^{\mathcal{E}}\big)\; P\big(T^{\mathcal{A}} \mid S^{\mathcal{E}}\big)\; dT^{\mathcal{A}} \tag{5}$$

一个两阶段过程生成文本标记（“内心独白”），然后是音频标记，将基于文本的推理与全双工能力结合起来。

预测同步。SyncLLM [20] 预测即将出现的用户段以最小化延迟：

$$\hat{e}_{t+1} \sim P\big(\cdot \mid S^{\mathcal{E}}_{\leq t}, S^{\mathcal{A}}_{\leq t}\big), \quad a_{t+1} \sim P\big(\cdot \mid S^{\mathcal{E}}_{\leq t}, \hat{e}_{t+1}, S^{\mathcal{A}}_{\leq t}\big) \tag{6}$$

3. 分类学

认知并行性，能够实现同时的语音编码和输出解码，需要脱离顺序 Transformer 架构。图 2显示了当前方法遵循的两种范式：通过模块化架构的**工程同步**和通过端到端系统的**学习同步**。

3.1. 工程同步

模块化方法通过专门的组件增强对话引擎，消除通过显式状态仲裁进行的重新训练。双工控制器——一种神经 FSM——扩展了声学 VAD 的功能，执行语义仲裁，区分中断、反馈通道和噪声。

外部控制器。 外部控制器保持与核心引擎的独立性。FlexDuo 引入了一个带有空闲状态的选择性注意力三元有限状态机 [7]。语义 VAD 使用轻量级（∼0.5B）模型分析 ASR 输出以最小化计算负载 [21]。VITA-1.5 在检测到中断时互换双实例的角色，通过增加计算成本来换取延迟 [22]。

内部控制器。 内部控制器将控制逻辑嵌入到引擎架构中。冻结-Omni 在冻结的 LLM 上按块进行状态预测 [23]; MinMo 的全双工预测器读取嵌入以做出让出回合的决策 [24]。神经-FSM 通过添加 FSM 令牌扩展词汇表，从而能够通过下一个标记预测实现自主状态管理 [8]。Mini-Omni2 通过语义状态令牌 [24] 实现基于命令的中断。

3.2. 学习同步

端到端架构原生处理双向音频流。继 dGSLM 从原始音频 [10] 展示出现的轮流对话后，这些系统使全双工能力成为固有特性。挑战在于协调 Transformer 的顺序性与对话的并行性。

模态接口。 模态界面在表示上有所不同。基于编解码器的方法 [10–12, 20, 25] 尽管序列变长，仍将音频离散化为令牌。SALMONN-omni 直接处理连续嵌入 [13]。SALM-Duplex 结合了连续输入和离散输出以实现准确性与延迟之间的权衡 [26]。

流处理。 流处理遵循多流或多路传输范式。多流方法如双塔架构使用交叉注意力进行同步 [10]，而 Moshi 的 RQ-Transformer 联合建模用户/代理音频和内部独白 [11]。单流方法将输入序列化为标准解码器：SyncLLM 使用同步令牌交错块 [20]，NTPP 使用成对因果掩蔽 [12]，LSLM/SALM-Duplex 探索不同的融合深度 [19, 26]。

Table 2. 跨四个维度对代表性开源 FD-SLM 进行综合评估。

模型	时间动力学		行为仲裁			语义连贯性		声学性能	
	脂肪组织 (↓)	样本量 (↓)	IRD (↓)	ISR (↑)	WER (↓)	PPL (↓)	问答准确率 (↑)	N-MOS (↑)	M-MOS (↑)
人类	∼0.20 s	∼0.30 s	2.32 s	93.69%	1.5%	10.2	92%	4.92 (±0.02)	4.85 (±0.03)
dGSLM	0.33 s (±0.12)	0.15 s (±0.03)	1.33 s	60.31%	25% (±3.4)	334.4	17.2%	3.85 (±0.12)	1.38 (±0.10)
NTPP	0.30 s (±0.15)	0.18 s (±0.05)	1.30 s	80.82%	7.5% (±1.22)	35	55.2%	4.15 (±0.06)	3.95 (±0.04)
如果	2.22 s (±0.70)	0.75 s (±0.10)	1.44 s	77.73%	5.20% (±0.13)	59.3	33.8%	3.90 (±0.07)	3.75 (±0.06)
萨尔蒙纳-全范围	0.38 s (±0.10)	0.25 s (±0.08)	1.38 s	85.6%	8.40% (±0.20)	21.1	61%	3.85 (±0.10)	3.95 (±0.15)
VITA-1.5	2.10 s (±0.65)	0.12 s (±0.05)	9.49 s	78.53%	5.45% (±0.10)	26.8	50.5%	4.00 (±0.08)	4.10 (±0.10)
冻结-万能	−0.40 s (±0.05)	1.11 s (±0.17)	9.25 s	54.97%	7.30% (±0.05)	30.2	56.9%	3.80 (±0.10)	3.90 (±0.07)

交互建模。 交互建模主要采用隐式动力学，其中模型通过沉默或生成可听见的令牌来控制轮流进行，而无需显式监督 [10–12,20,25]。相比之下，SALMONN-omni 的动态思维机制 [13] 生成控制令牌以实现显式的状态管理，将大语言模型定位为端到端框架中的双向预测器。

4. 评估

FD-SLMs 需要在三个相互依存的轴上进行协调评估：流架构实现实时交互、对话训练数据以及全面的基准测试方法。

4.1. 架构组件

FD-SLMs 需要专门的流架构以实现低于-200 毫秒的延迟，从而支持自然交替发言。[14,15] 表 1总结了三个关键阶段中的策略。

输入感知。 具有最小前瞻性的连续编码是必不可少的。虽然传统编解码器需要因果适应，专门设计的流式编解码器原生运行 [13,27,28]。离散范例采用严格的因果/近零前瞻性神经编解码器 [11,12]；标记化块粒度根本上限制了感知延迟 [12,29–31]。

核心处理。 并发流通过交叉注意力 [10]、联合自回归 [11]、预测同步 [20] 或显式控制机制 [7,24] 进行同步。一个 100 – 200 毫秒的“认知时钟”设定感知—反应粒度 [9,11,20]。KV-缓存效率直接影响持续响应性 [12,32]。

输出综合。 离散模型重复使用编解码器解码器以实现最小延迟 [33,34]。连续流水线采用分块感知流匹配 [35]、固定长度交错生成 [13] 或紧密耦合的大语言模型 – 声码器堆栈 [22]。

4.2. 训练数据

数据稀缺仍然是关键问题：FD-SLMs 需要同步多通道自发对话，而单声道语料库无法提供这一点。当前的训练使用了有限的数据集 [10–12]，限制了多样性（请参见表 3的示例；完整列表请参阅我们的仓库）。

Table 3. 公开可用的数据集用于 FD-SLM

Dataset	Lang	Scene	Channels	Hours
AMI Meeting Corpus	EN	meeting	8	100
ICSI Meeting Corpus	EN	meeting	6	70
LibriCSS	EN	meeting	7	10
Fisher English	EN	phone	2	1,960
SEAME (Mandarin–English CS)	EN+ZH	interview	2	192
HKUST Mandarin Telephone	ZH	phone	2	149

合成 TTS 生成 [20] 未能捕捉到韵律同步和重叠动态，限制了泛化能力。进步需要端到端的对话合成以及单通道数据的高级源分离技术。

4.3. 基准框架

传统上为半双工系统设计的指标 [1,36] 无法捕捉实时全双工行为：当模型发言时，它们如何介入以及会话控制仲裁 [6]。

通过模型特定的指标 [8,10–12] 进行的历史分割阻碍了系统的比较。最近的标准化努力 [16–18] 使我们能够通过四支柱分类法（图 3）实现可重复评估。

表 2揭示了关键差距：虽然音质接近人类水平，但时间动态变化广泛，行为仲裁表现不佳（ISR: 54-86% 对比 94% 人类），且语义连贯性与响应性之间存

Fig. 3. 四个-支柱基准测试 FD-SLMs 的分类法。

在权衡——表明达到与人类持平的 FD 需要范式的架构进步。

5. 结论

FD-SLMs 标志着从回合制到同步对话的范式转变。通过认知并发形式化以及我们区分工程同步和学习同步的分类法，我们阐明了基本的设计权衡。我们的四支柱评估显示，尽管声学质量接近人类水平，但仍存在关键差距：不一致的时间动态、次优的行为仲裁和反比的延迟-连贯性相关性。

进步需要解决相互关联的挑战。架构碎片化阻碍了与大语言模型扩展规律相一致的大规模设计。数据匮乏——特别是同步多通道记录和非英语资源 [37]——限制了学习。当前评估缺乏主动行为指标 [25]，而超低延迟引入了需要实时过滤的安全风险。

推进 FD-SLMs 需要架构收敛、合成数据捕捉真实动态、全面的行为评估和强大的安全机制。只有通过协调努力，我们才能实现真正类人的对话 AI，这种 AI 是响应式的、可扩展的，并且可以道德地部署。

6. REFERENCES

[1] OpenAI et al., "GPT-4o system card," 2024.

[2] Boyong Wu et al., "Step-audio 2 technical report," 2025.

[3] Tanya Stivers et al., "Universals and cultural variation in turn-taking in conversation," PNAS, 2009.

[4] Stephen C. Levinson and Francisco Torreira, "Timing in turn-taking and its implications for processing models of language," Frontiers in Psychology, 2015.

[5] Antoine Raux and Maxine Eskenazi, "A finite-state turn-taking model for spoken dialog systems," in Proc. NAACL-HLT, 2009.

[6] Gabriel Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," Computer Speech & Language, 2021.

[7] Ziyang Liao et al., "Flexduo: A pluggable system for enhancing spoken dialogue models with full-duplex capabilities," 2025.

[8] Zheng Wang et al., "Neural-FSM: A full-duplex speech dialogue scheme based on large language model," in Proc. NeurIPS, 2024.

[9] Jun Zhang et al., "Omniflatten: A unified framework for spoken language model via progressive flattening," in Proc. ACL, 2025.

[10] Anh-Duy Nguyen et al., "Generative spoken dialogue language modeling," TACL, 2023.

[11] Alexandre Défossez et al., "Moshi: a speech-text foundation model for real-time dialogue," 2024.

[12] Qichao Wang et al., "NTPP: Generative speech language modeling for dual-channel spoken dialogue via next-token-pair prediction," 2025.

[13] Dong Yu et al., "Salmonn-omni: A codec-free LLM for full-duplex speech understanding and generation," 2025.

[14] Siddhant Arora et al., "On the landscape of spoken language models: A comprehensive survey," 2025.

[15] Jia Cui et al., "Recent advances in speech language models: A survey," 2025.

[16] Yizhou Peng et al., "Fd-bench: A full-duplex benchmarking pipeline designed for full duplex spoken dialogue systems," 2025.

[17] Guan-Ting Lin et al., "Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities," 2025.

[18] Guan-Ting Lin et al., "Full-duplex-bench v1.5: Evaluating overlap handling for full-duplex speech models," 2025.

[19] Ziyang Ma et al., "Language model can listen while speaking," 2024.

[20] Aditya Veluri et al., "Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents," 2024.

[21] Mohan Shi et al., "Semantic VAD: Low-latency voice activity detection for speech interaction," 2023.

[22] Chaoyou Fu et al., "Vita-1.5: Towards GPT-4o level real-time vision and speech interaction," 2025.

[23] Xiong Wang et al., "Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM," 2024.

[24] Zhifei Xie et al., "Mini-omni2: Towards open-source GPT-4o with vision, speech and duplex capabilities," 2024.

[25] Yemin Shi et al., "Voila: A sophisticated, synchronous, and swift spoken language model," 2025.

[26] Ke Hu et al., "Salm-duplex: Efficient and direct duplex modeling for speech-to-speech language model," 2025.

[27] Yangyang Shi et al., "Emformer: Efficient memory transformer based acoustic model for low latency streaming speech recognition," in Proc. ICASSP, 2021.

[28] Zengwei Yao et al., "Zipformer: A faster and

better encoder for automatic speech recognition," in Proc. ICLR, 2024.

[29] Pooneh Mousavi et al., "Discrete audio tokens: More than a survey!," 2025.

[30] Rithesh Kumar et al., "High-fidelity audio compression with improved RVQGAN," in Proc. NeurIPS, 2023.

[31] Yuanzhe Xu et al., "Wavtokenizer: A novel general-purpose audio-to-token converter," in Proc. ICLR, 2025.

[32] Zhen Li et al., "Connector-s: A survey of connectors in multimodal large language models," 2025.

[33] Zhengyan Sheng et al., "Syncspeech: Low-latency and efficient dual-stream text-to-speech based on temporal masked transformer," 2025.

[34] Sambal Shikhar et al., "LLMVoX: A zero-shot, personalized, and streaming speech synthesis leveraging large language models," in Findings of ACL, 2025.

[35] Qian Chen et al., "Minmo: A multimodal large language model for seamless voice interaction," 2025.

[36] Takaaki Saeki et al., "UTMOS: UTokyo-sarulab MOS prediction system for voice conversion challenge 2022," in Proc. SSW, 2022.

[37] Shintaro Ohashi et al., "Towards a japanese full-duplex spoken dialogue system: Data collection, modeling, and evaluation," 2025.

[38] OpenBMB et al., "Minicpm-llama3-v 2.5: An 8b-scale multimodal LLM," 2024.

[39] Jing Peng et al., "A survey on speech large language models for understanding," 2024.

[40] Jinglin Chen et al., "Wavchat: A survey of spoken dialogue models," 2024.

[41] Guillermo Castillo-López et al., "A survey of recent advances on turn-taking modeling in spoken dialogue systems," in Proc. IWSDS, 2025.

[42] Alexandre Défossez et al., "High-fidelity neural audio compression," Trans. Mach. Learn. Res., 2023.

[43] Hubert Siuzdak, "Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis," in Proc. ICLR, 2024.