arXiv:2603.08179v1 [eess.AS] 9 Mar 2026

隐私保护的端到端全双工语音对话模型

Nikita Kuzmin[1,2], *Tao Zhong*[3,4], *Jiajun Deng*[3,†], *Yingke Zhu*[3], *Tristan Tsoi*[3], *Tianxiang Cao*[3], *Simon Lui*[3], *Kong Aik Lee*[5], *Eng Siong Chng*[1]

[1]新加坡南洋理工大学 [2]新加坡科技研究局，新加坡[3]华为莱布尼茨研究中心 [4]香港中文大学[5]香港理工大学s220028@e.ntu.edu.sg

摘要

端到端全双工语音模型通过始终在线的大语言模型主干处理用户音频，但其隐藏表征对说话人隐私的影响尚未得到研究。遵循 VoicePrivacy 2024 协议并采用懒惰知情攻击者，我们发现 SALM-Duplex 和 Moshi 的隐藏状态在所有 Transformer 层中都会泄露大量说话人身份信息。逐层和逐轮分析表明，泄露在所有层中持续存在：SALM-Duplex 在早期层中的泄露更强，而 Moshi 的泄露则较为均匀；可链接性会在最初几轮内急剧上升。我们提出两种基于 Stream-Voice-Anon 的流式匿名化设置：波形级前端（Anon-W2W）和特征域替换（Anon-W2F）。与离散编码器基线相比，Anon-W2F 将等错误率提高了 3.5×以上（11.2% →41.0%），接近 50% 的随机猜测上限；而在各项设置下，Anon-W2W 保留了基线 sBERT 性能的 78–93%，响应时延低于一秒（FRL 小于 0.8 秒）。[1]

索引词：说话人匿名化、全双工语音、隐私、说话人验证、语音智能体

1. 引言

端到端（E2E）全双工语音对话系统代表了从轮流交互到始终开启、同时聆听与说话的根本转变。SALM-Duplex [1], Moshi [2], 和 SyncLLM [3] 等模型通过仅解码器大语言模型骨干网络持续处理原始用户音频，在整个对话过程中于每个 Transformer 层计算隐藏状态表示。不同于分别处理用户流和智能体流的级联系统，这些全双工系统的大语言模型核心会在用户语音流上维持持久的内部状态，从而捕捉用户的声音、说话风格和身份。

这造成了一种迄今鲜少探索的隐私暴露风险。根据《通用数据保护条例》及类似法规，即使没有被外部访问利用，模型表征中仅仅存在可识别的说话人信息，也会构成合规风险，因此必须主动审计并缓解此类信息泄露。针对自监督语音模型（如 wav2vec 2.0、HuBERT、WavLM）的探测研究已证实，其隐藏表征强烈编码了说话人身份 [4, 5, 6]。在文本领域，既有研究表明，大语言模型隐藏状态可能泄露用户的人口统计特征和内容层面的属性 [7], ，而成员推断探测器也揭示了掩码语言模型中的隐私风险 [8]。我们的研究重点有所不同：我们探讨的是——

[†]表示通讯作者。[1]演示页面: https://anonymous-569230593.github. io/

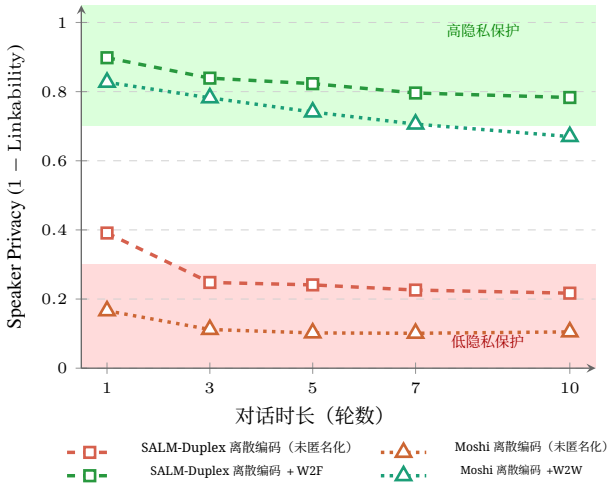


图1： 说话人隐私（1 − 可链接性）与对话轮数的关系。图中仅展示离散编码器变体；为清晰起见，未展示连续编码器的可链接性。红线：未进行匿名化；绿线：匿名化后。未进行匿名化时，两个系统在几轮对话内都会跌入低隐私区域；匿名化可将隐私提升至受保护区域，但 *Moshi + W2W* 会随着对话时长增加而逐渐退化。

我们研究说话人身份泄露，即对说话者是谁的持久编码；无论对话内容如何，这种编码都可能导致重新识别。然而，据我们所知，尚无既有研究考察这些始终开启的大语言模型隐藏状态是否保留了足以实现重新识别的说话人身份信息。

我们通过实证研究，回答了这一问题，研究对象是两种颇具代表性的端到端全双工系统：SALM-Duplex 和 Moshi [2]。我们遵循 VoicePrivacy 2024 挑战赛 [9] 评估协议，采用惰性知情攻击者场景，在从各层组提取的隐藏状态表示上训练说话人验证攻击者（探测器）。等错误率（EER）是我们的主要隐私度量指标；此外，我们还报告了 [10]经法律验证的评估框架所采用的可链接性指标。原始 SALM-Duplex（连续编码器）的平均池化 EER 为 28.5%，表明说话人身份泄露严重；而 Moshi 的离散编码器达到 6.4% 的 EER（几乎可以完美识别）。我们还提出并评估了两种流式匿名化设置（图 2），采用 Stream-Voice-Anon [11]，以缓解这种泄露，同时不牺牲对话效用。

我们的贡献如下：

• 我们分析了两个端到端全双工对话大语言模型（SALM-Duplex 和

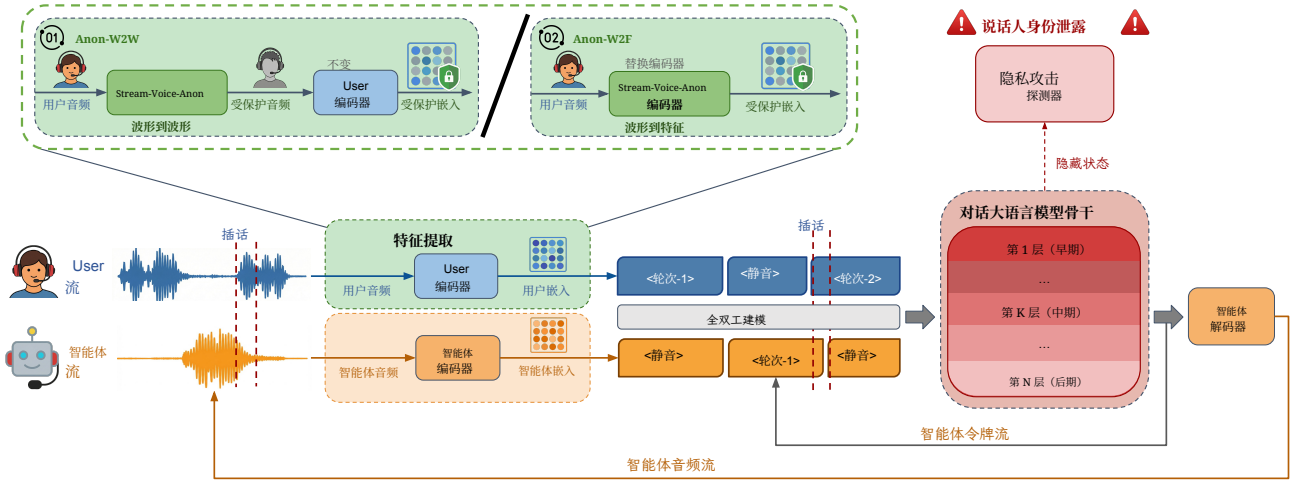


图 2： 原始 *SALM-Duplex* 流程与所提出的匿名化设置概览。主图展示了基于自动语音识别的编码器基线：连接到大语言模型隐藏状态的 *ECAPA-TDNN* 探测器（红色虚线路径）揭示出大量说话人身份泄露。左上角的 *Anon-W2W* 插图在未经修改的自动语音识别编码器之前加入 *Stream-Voice-Anon*，以对波形进行匿名化。右上角的 *Anon-W2F* 插图则以 *Stream-Voice-Anon* 编码器替代自动语音识别编码器（启用匿名化），并对大语言模型进行微调，从而消除冗余的波形合成步骤。（*Anon-W2F* 针对 *SALM-Duplex* 展示；*Anon-W2W* 设置还在 *Moshi* 上进行了评估。）

Moshi）隐藏状态中的说话人身份泄露，并将探测方法 [12, 4] 从静态编码器扩展至始终运行的对话骨干网络。

- 我们提供逐层分析和按轮次长度分析，展示大语言模型的哪些部分承载了最多的说话人身份信息，以及信息泄露如何随着对话长度增加而累积。
- 我们提出两种流式匿名化设置：Anon-W2W 在波形级应用 Stream-Voice-Anon （保留原有编码器，并在 SALM-Duplex 和 Moshi 上均进行了验证）；Anon-W2F 则以离散编码器替换连续编码器，并在特征级启用匿名化。

2. 相关工作

2.1. 端到端语音对话语言模型

生成式口语语言模型（GSLM） [13] 确立了直接基于离散单元对语音建模的范式，随后 dGSLM [14], 将其扩展至双通道口语对话，成为首个完全端到端的全双工对话模型。SpeechGPT [15] 和 SpiRit-LM [16] 证明，单一大语言模型主干可以原生交错处理语音和文本词元，以实现对话交互。SyncLLM [3] 引入了同步的大语言模型推理，以支持实时全双工轮次管理。最新的系统 SALM-Duplex [1] 和 Moshi [2], 通过连续双流处理实现了低延迟、始终在线的运行。关键在于，由于这些模型都将用户原始语音作为输入，说话人身份信息可能会在无意间编码到其大语言模型的隐藏表示中，但迄今为止，尚无研究从隐私视角对这一问题进行分析。

2.2. 说话人匿名化

VoicePrivacy 挑战赛 [17, 9] 建立了用于说话人匿名化的标准化评估协议和基准，以自动说话人验证系统的等错误率作为主要隐私指标；Vauquier 等人 [10] 最近在此基础上进行了扩展，加入了一个经过法律验证的框架，其中纳入了符合《通用数据保护条例》要求的可链接性指标。匿名化方法包括基于 x-vector 的方法 [18, 17], 、扰动 [19],语音中间表征 [20], 、神经音频编解码器语言模型 [21, 22], 以及解耦语音表征 [23, 24]。包括 Stream-Voice-Anon [11], 、TVTSyn [25], 、DarkStream [26], 和 [27], 在内的流式匿名化系统，将这些方法扩展到了端到端双工模型所需的实时场景；鉴于其具有竞争力的隐私与效用权衡以及开源可用性[2]，我们选择 Stream-Voice-Anon 作为前端。所有这些方法都在波形级上运行。就我们所知，我们的工作首次分析并缓解了全双工大语言模型隐藏状态内部的说话人身份泄露。

2.3. 神经语音与语言表征中的隐私

探测研究已证实，自监督语音模型在其 Transformer 层中编码了大量说话人身份信息：Pasad 等人 [4] 发现，说话人信息在 wav2vec 2.0 的较低层和中间层达到峰值；SUPERB [5] 则证明，轻量级探测器可基于冻结的自监督学习表征实现高效的说话人验证。Chiu 等人 [6] 最近在多种自监督学习架构和说话人属性上进行了大规模研究，进一步证实了这一点。Tenney 等人 [12] 则在文本模型中确立了逐层探测方法。我们的研究重点有所不同：我们不保护对话内容，而是研究说话人身份泄露，即对说话者身份的持续编码；这使得无论说话内容是什么，都可能重新识别说话人，并且还可能泄露性别、口音和健康状况等属性。Nautsch 等人 [28] 全面综述了说话人与语音特征描述中的隐私威胁，并指出从音频表征中重新识别说话人是一类主要风险。

我们将这一研究方向扩展至端到端全双工对话大语言模型。与静态自监督学习编码器不同，这些模型的隐藏状态会在实时用户语音上连续计算。

[2]https://github.com/Plachtaa/StreamVoiceAnon

表 1： *VPC2024* 评测集上的隐私、对话质量和效率。隐私：等错误率越高、可链接性（*Lnk*）越低 = 隐私保护越强；等错误率为 = *50%* = 表示实现完美匿名化。未匿名化行的可链接性使用 *"orig"* 攻击者；匿名化行则使用 *"lazy-informed"* 攻击者 *[10]*。质量：*sBLEU* / *sBERT* 越高 = 效果越好。效率：*RTFx* = *1/RTF*；*FRL* = 首次响应延迟（秒）；*TTSR* = 轮流发言成功率；*Int.L.* = 打断延迟（秒）；*ISR* = 打断成功率。粗体 = 表示每列最佳值；下划线 = 表示次佳值。阴影行 = 表示匿名化条件。

模型 Name	User 编码器	匿名化 Type	隐私		质量				效率				
			等错误率↑	可链接性↓	sBLEU↑ S2T	sBLEU↑ S2S	sBERT↑ S2T	sBERT↑ S2S	实时因子↑	FRL↓	TTSR↑	Int.L.↓	ISR↑
Moshi	离散	–	6.4	0.90	**7.18**	**6.85**	0.50	0.48	17	0.50	0.85	1.24	0.45
	离散	W2W	36.9	0.35	5.20		0.39	0.38	1.6	0.72	0.72	1.31	0.38
SALM-双工	离散	–	11.2	0.79	4.18	3.15	0.45	0.29	238	**0.44**	0.85	1.03	0.96
	离散	W2F	**41.0**	**0.23**	3.57	1.49	0.39	0.20	2.5	0.50	0.68	1.18	0.93
	连续	–	28.5	0.29	6.91	6.35	**0.59**	**0.50**	**263**	0.68	**0.97**	0.60	**0.99**
	连续	W2W	34.6	0.24	6.46				1.7	0.80	0.94	**0.57**	0.98

3. 方法

3.1. 端到端全双工对话模型

我们分析了两个端到端全双工对话系统。**Moshi** [2]采用仅解码器 Transformer 主干网络，并结合残差量化（RVQ）音频编解码器编码器，在单个自回归过程中联合建模用户流和系统流。**SALM-Duplex** [1] 是一个仅解码器大语言模型，通过经自动语音识别初始化的连续编码器对用户音频进行编码，并以帧同步方式处理用户音频流和智能体音频流，在整个对话过程中维持持久的隐藏状态；我们重新实现了一个修改版本，采用了来自 [29] 的语音解码器架构，但不包含时间顺序思考机制。我们的重新实现取得了相当的响应质量；时间顺序思考机制对隐私的影响留待未来研究。两个系统都会在连续用户语音上维持持久的大语言模型隐藏状态，因此适合进行说话人身份探测。

3.2. 用户编码器变体

我们根据输出表示将编码器分为两类，这决定了有多少说话人身份信息会传递到大语言模型骨干网络。

离散编码器（Moshi 和 SALM-Duplex 变体）：Moshi 原生使用 RVQ 编解码器编码器，生成离散的词元表示。对于 SALM-Duplex，我们将连续前端替换为基于 Firefly 架构的离散编码器（实现细节见第 3.5 节）。由于这两种离散编码器均经过高保真语音重建训练，因此它们的词元表示保留了丰富的说话人信息。SALM-Duplex 离散编码器可在两种模式下运行：匿名化禁用，作为我们的消融实验，以单独考察编码器替换的影响；或通过 Stream-Voice-Anon [11],启用匿名化，构成我们的 Anon-W2F 设置（第 3.3 节）。

连续编码器（原始 SALM-Duplex）：一个经过自动语音识别初始化的适配器将原始用户音频编码为连续嵌入。由于自动语音识别预训练优化的是语音内容，而非说话人特征，因此，正如我们在第 4 节中所展示的，该编码器是内置部分隐私保护的一种良好设计选择，不过其隐私上限有限。

3.3. 匿名化设置

我们提出了两种使用 Stream-Voice-Anon [11] 的流式匿名化设置。

3.3.1. Anon-W2W：波形到波形匿名化

Anon-W2W 将连续编码器与 SALM-Duplex 配对，并将 Moshi 的原生离散编解码器编码器用于处理。Stream-Voice-Anon 作为预处理步骤，将用户原始波形转换为匿名化波形，再输入对话模型。然而，这会引入冗余的处理步骤：匿名化波形先经过合成，然后再由模型原有的编码器重新编码。

3.3.2. Anon-W2F：波形到特征匿名化

Anon-W2F 与离散编码器配对（启用匿名化）。我们将 SALM-Duplex 的连续编码器前端替换为离散编码器，并启用 Stream-Voice-Anon 的匿名化模块，该模块原生处理离散令牌表示。由于匿名化步骤在与编码器相同的特征域中执行，因此无需 Anon-W2W 中冗余的波形合成步骤。修改后的模型使用与 [29] 相似的数据组合进行预训练（约 12k 小时的多轮对话和 2.7k 小时的问答数据），随后在 InstructS2S-200K[3] [30] 上进行微调。此设置需要进行架构修改，但通过原生于特征域的匿名化实现了更强的隐私保障。尽管 Anon-W2F 的原理（将模型的编码器替换为支持匿名化的离散编码器）在架构上可迁移，但其效果取决于目标模型的分词器兼容性；我们在 SALM-Duplex 上展示了这一方法，并将 Moshi 集成留待后续工作。

3.4. 隐藏状态提取

我们从 Moshi（N=32 个Transformer层）和 SALM-Duplex（N=20）中提取隐藏状态表示，分别取三个单独的层：**早期**（第1层）、**中期**（第 $N/2$层）和**后期**（第 N层），以及一个**均值池化** 表示，即对全部 N 层进行平均（在表2中记为"All"）。时序

[3]https://huggingface.co/datasets/ICTNLP/ InstructS2S-200K

表2：说话人验证攻击者的逐层等错误率（%）。 = 越高，隐私保护效果越好；*50%* = 为随机水平。所有数值均为对所有层进行 = 均值池化后的结果。粗体 = 表示最佳；下划线 = 表示次佳。阴影行 =表示匿名化条件。

系统	编码器	匿名化	早期	中期	后期		All
Moshi	离散	–	7.3	5.6	6.4	6.4	
	离散	W2W	42.5	37.6	___	___	
SALM-双工	离散	–	7.5	14.0	20.1	11.2	
	离散	W2F	**43.8**	**40.5**			
	连续	–	24.6	28.6	32.1	28.5	
	连续	W2W	31.5	33.7			

池化由探针架构内部处理。

3.5. 设置

数据集。隐私度量指标在 VoicePrivacy 2024 挑战赛 [9] 评测集上进行评估，该评测集源自 LibriSpeech 的 dev-clean 和 test-clean 划分。尽管该数据集由朗读语音而非即兴对话构成，但鉴于其标准化的评估协议，它仍是一个良好的起点；在对话语料库上进行评估是未来工作的一个重要方向。说话人验证攻击者在 LibriSpeech train-clean-360 上训练。质量和效率指标在 MtBenchEval [31, 32], 上进行评估，这是由 URO-Bench 标准化的多轮对话基准。Moshi 使用开源预训练权重；SALM-Duplex 按照 [29] 的语音解码器架构重新实现。

说话人验证攻击者。我们分别从头开始训练用于 SALM-Duplex 和 Moshi 的 ECAPA-时延神经网络 [33] 攻击者，复现 VPC2024 [9] 中的惰性知情攻击者场景：采用相同的超参数，并使用适配各模型隐藏状态维度的输入投影层。

隐私度量指标。遵循 VoicePrivacy 2024 挑战赛 [9] 的协议，我们将等错误率（EER%）作为主要隐私度量指标。EER 越高，表示隐私保护越强；EER $\approx$ 50% 对应于随机猜测的判别水平，即完全匿名化。为进一步加强隐私分析，我们还报告可链接性 [10]。

效用度量指标。我们使用根据 URO-Bench [31], 全双工基准 [32], 和 SALM-Duplex [1]改编的评估脚本来评估对话质量。质量指标（sBLEU、sBERT）针对语音转文本（S2T）和语音转语音（S2S）响应模式分别报告：在 S2T 模式下，模型直接输出文本；在 S2S 模式下，模型输出语音，随后由自动语音识别系统转写。效率指标包括实时因子加速比（RTFx）、首次响应延迟（FRL）、轮流发言成功率（TTSR）、打断延迟（Int.L.）和打断成功率（ISR），并按系统条件分别报告（表 1）。所有 RTFx 和延迟测量值均在单个图形处理器上估算。

4. 结果

表 1 报告了隐私、响应质量和效率指标。为实现高保真语音重建而训练的离散编码器，比连续编码器泄露了更多说话人身份信息（Moshi：等错误率为 6.4%；SALM-Duplex 离散编码器：11.2%，连续编码器：28.5%），这证实自动语音识别预训练是一种有利于隐私的设计选择，因为它会舍弃说话人特征，转而保留语言内容。两种匿名化设置都显著减少了信息泄露：Anon-W2W 将 Moshi 的等错误率提高至 36.9%（增加 +30.5 个百分点），将 SALM-Duplex 连续编码器的等错误率提高至 34.6%（增加 +6.1 个百分点）；而 Anon-W2F 达到 41.0% 的等错误率，接近 50% 的随机猜测上限。关键的是，将 Anon-W2F 与仅使用离散编码器的消融实验（11.2%）进行比较，证实这一提升完全来自匿名化，而非编码器替换。这些规律在两种架构中均成立，证实说话人身份暴露是端到端全双工大语言模型隐藏状态的一项稳定属性。我们的 ECAPA-TDNN 探测器给出了信息泄露的下界；VoicePrivacy 攻击者挑战赛 [34] 表明，更复杂的攻击者能够进一步降低等错误率，这意味着实际隐私风险可能高于本文报告的水平。

匿名化会导致中等程度的质量下降（sBERT S2T 相对下降 7–22%），但隐私收益始终远大于这一代价，各种设置下的等错误率相对改善了 21–477%。我们注意到，基于文本的质量指标无法反映自然度和韵律等语音级属性；语音质量评估（例如平均意见分、UTMOS）仍有待未来研究。加入匿名化模块会显著降低实时因子（从 17–263$\times$ 降至 1.6–2.5$\times$），因为匿名化器占用了大部分推理时间，但所有条件仍能满足实时运行要求（实时因子 > 1）。Anon-W2F 比 Anon-W2W 更快（实时因子分别为 2.5 和 1.6–1.7），这是因为特征域编码器避免了冗余的波形合成步骤。降低匿名化模块的计算成本是未来研究的一个重要方向。

4.1. 消融研究

逐层分析。表2展示了各层的等错误率。Moshi 在所有层的等错误率均较低（5.6–7.3%），而两个 SALM-Duplex 变体则呈现出从早期层到后期层泄露逐渐减少的趋势，这与更深层逐步抽象掉说话人特征的情况相符。匿名化使所有层组的等错误率均有所提高，所有匿名化条件下的数值均达到 31–44%。

轮次长度分析。图1绘制了 $1 -$ 可链接性与轮次数的关系。未进行匿名化时，隐私保护能力在最初几轮内迅速下降。即使经过 10 轮，两种匿名化系统仍能维持可接受的保护水平（SALM-Duplex + W2F：0.90$\rightarrow$0.78；Moshi +W2W：0.83$\rightarrow$0.67）。

5. 结论

我们证明，显著的说话人身份泄露是端到端全双工架构的一项普遍特性，这一点在 SALM-Duplex 和 Moshi 中均得到验证；与经过 ASR 预训练的连续编码器相比，离散编码器泄露的信息多得多。波形级匿名化（Anon-W2W）能够降低这两种架构中的风险，其中相对收益最大的是暴露程度最高的系统；而特征域匿名化（Anon-W2F）提供了最强保护，使等错误率提高 3.5 倍以上，并接近随机水平。这些发现表明，常开语音人工智能系统应贯彻隐私保护设计。

未来的工作包括将 Anon-W2F 扩展到 Moshi 和其他端到端全双工架构，研究如何在最大限度减少隐私威胁的同时实现个性化，改进隐私保护措施以降低其对响应质量和延迟的影响，并在更强大、更多样化的攻击者模型下评估隐私。

6. 生成式人工智能使用披露

研究人员使用生成式人工智能工具进行了校对和语法纠正，并做了少量修改。所有研究构想、实验设计、实施和分析均由作者完成。

7. 参考文献

[1] 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”，载于《*Interspeech 2025* 论文集》，2025年，第2715–2719页。[2] 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，arXiv预印本 arXiv:2410.00037。[3] B. Veluri, B. N. Peloquin, B. Yu, H. Gong, and S. Gollakota, “超越轮次式接口：同步大语言模型作为全双工对话智能体”，载于《*2024*年自然语言处理实证方法会议论文集》（*EMNLP*）。计算语言学协会，2024年。[4] A. Pasad, J.-C. Chou, and K. Livescu, “自监督语音表征模型的逐层分析”，载于《*IEEE* 自动语音识别与理解研讨会》（*ASRU*）。IEEE，2021年，第914–921页。[5] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.- W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, “SUPERB：语音处理通用性能基准”，载于《*Interspeech 2021* 论文集》，2021年。[6] A. Y. F. Chiu, K. C. Fung, R. T. Y. Li, J. Li, and T. Lee, “自监督语音表征中说话人特定属性的大规模探测分析”，2025年，arXiv预印本 arXiv:2501.05310。[7] R. Staab, M. Vero, M. Balunovi´c, and M. Vechev, “超越记忆：通过大语言模型推断侵犯隐私”，载于《第*12*届国际学习表征会议论文集》（*ICLR*），2024年。[8] F. Mireshghallah, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, a nd R. Shokri, “使用成员推理攻击量化掩码语言模型的隐私风险”，载于《自然语言处理实证方法会议论文集》（*EMNLP*），2022年。[9] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. W. D. Evans, J. Yamagishi, and M. Todisco, “VoicePrivacy 2024 挑战赛评估计划”，2024年，arXiv预印本 arXiv:2404.02677。[10] N. Vauquier, B. M. L. Srivastava, S. A. Hosseini, and E. Vincent, “经法律验证的语音匿名化评估框架”，载于《*Interspeech 2025* 论文集》，2025年。[11] N. Kuzmin, S. Liu, K. A. Lee, and E. S. Chng, “Stream-Voice-Anon：通过神经音频编解码器和语言模型提升实时说话人匿名化的效用”，载于《*IEEE* 声学、语音与信号处理国际会议论文集》（*ICASSP*），2026年。[12] I. Tenney, D. Das, and E. Pavlick, “BERT 重新发现经典自然语言处理流程”，载于《第*57*届计算语言学协会年会论文集》（*ACL*），2019年。[13] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T. A. Nguyen, J. Copet, A. Baeveski, A. Mohamed, and E. Dupoux, “基于原始音频的生成式口语语言建模”，刊于《计算语言学协会汇刊》，第9卷，第1336–1354页，2021年。[14] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed, and E. Dupoux, “生成式口语对话语言建模”，刊于《计算语言学协会汇刊》，第11卷，第250–266页，2023年。

[15] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, 和 X. Qiu，“SpeechGPT：赋予大语言模型内在的跨模态对话能力”，载于计算语言学协会：*EMNLP 2023*论文集，2023年。[16] T. A. Nguyen, B. Muller, B. Yu, M. R. Costa-juss`a, M. Elbayad, S. Popuri, P.-A. Duquenne, R. Algayres, R. Mavlyutov, I. Gat, G. Synnaeve, J. Pino, B. Sagot, 和 E. Dupoux，“SpiRit-LM：交错式口语和书面语语言模型”，计算语言学协会会刊，2024年。[17] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. W. D. Evans, E. Vincent, 和 J. Yam- agishi，“VoicePrivacy 2022 挑战赛：语音匿名化的进展与展望”，*IEEE/ACM* 音频、语音与语言处理汇刊，第32卷，第3477–3491页，2024年。[18] S. Meyer, F. Lux, P. Denisov, J. Koch, P. Jaubert, 和 N. T. Vu，“利用生成对抗网络匿名化语音以保护说话人隐私”，载于*IEEE* 口语语言技术研讨会（*SLT*）论文集，2022年。[19] N. Kuzmin, H.-T. Luong, K. A. Lee, 和 E. S. Chng，“VoicePrivacy 2024 挑战赛的 NTU-NPU 系统”，载于第四届语音通信安全与隐私研讨会（*SPSC*）论文集，2024年。[20] S. Meyer 等，“基于语音中间表示的说话人匿名化”，载于*Interspeech 2022* 论文集，韩国仁川，2022年。[21] M. Panariello, F. Nespoli, M. Todisco, 和 N. Evans，“利用神经音频编解码器语言模型进行说话人匿名化”，载于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*）论文集，2024年。[22] X. Yao, N. Kuzmin, K. A. Lee, 和 E. S. Chng，“VoicePrivacy 2024 挑战赛的 NPU-NTU 系统”，载于第四届语音通信安全与隐私研讨会（*SPSC*）论文集，2024年。[23] P. Champion, D. Jouvet, 和 A. Larcher，“构建语音匿名化系统是否只需解耦表示？”，载于*Interspeech 2022* 论文集，韩国仁川，2022年。[24] R. Aloufi, H. Haddadi, 和 D. Boyle，“通过解耦表示实现隐私保护语音分析”，载于*2020* 年 *ACM SIGSAC* 云计算安全会议研讨会（*CCSW*）论文集，2020年。[25] W. Quamer, M.-R. Tseng, G. Nasrallah, 和 R. Gutierrez-Osuna，“TVTSyn：用于流式语音转换和匿名化的内容同步时变音色”，载于第十四届国际学习表征会议，2026年。[26] W. Quamer 和 R. Gutierrez-Osuna，“DarkStream：低延迟实时语音匿名化”，载于*IEEE* 自动语音识别与理解研讨会（*ASRU*）论文集，2025年。[27] W. Quamer 和 R. Gutierrez-Osuna，“用于低延迟语音匿名化的端到端流式模型”，载于*IEEE* 口语语言技术研讨会（*SLT*）。IEEE，2024年，第727–734页。[28] A. Nautsch, A. Jim´enez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delanne, M. Todisco, T. Schneider, 和 N. W. D. Evans，“保护说话人和语音特征描述中的隐私”，《计算机语音与语言》，第58卷，第441–480页，2019年。[29] Z. Wu 等，“全双工口语对话语言模型中的时序思维”，2025年，arXiv预印本 arXiv:2510.05150。[30] Q. Fang 等，“LLaMA-Omni 2：基于大语言模型的实时语音聊天机器人”，载于第*63*届计算语言学协会年会（*ACL*）论文集，2025年。[31] Yan 等，“URO-Bench：端到端口语对话模型综合基准测试”，2025年，arXiv预印本 arXiv:2502.17810。[32] Lin 等，“Full-duplex-bench v1.5：全双工口语对话模型基准测试”，2025年，arXiv预印本 arXiv:2507.23159。

[33] B. Desplanques, J. Thienpondt, 和 K. Demuynck, “ECAPA-TDNN：基于 TDNN 的说话人验证中的强调通道注意力、传播与聚合”，载于*Interspeech 2020* 论文集，中国上海，2020年，第3830–3834页。[34] N. Tomashenko, X. Miao, E. Vincent, 和 J. Yamagishi, “VoicePrivacy 2024 攻击者子挑战赛”，载于*IEEE* 声学、语音与信号处理国际会议（*ICASSP*）论文集，2025年。