arXiv:2603.06079v1 [eess.AS] 6 Mar 2026

StreamVoiceAnon+：通过帧级声学蒸馏实现情感保留流式说话人匿名化

Nikita Kuzmin[1,2], *Kong Aik Lee*[3], *Eng Siong Chng*[1]

[1]新加坡南洋理工大学[2]新加坡科技研究局（A*STAR）资讯通信研究院（I2R）[3]香港香港理工大学

s220028@e.ntu.edu.sg

摘要

我们致力于解决流式说话人匿名化（SA）中保留情感内容的难题。为音频续接训练的神经音频编解码器语言模型往往会削弱源情绪：内容令牌会丢弃情感信息，模型也会默认采用占主导地位的声学模式，而非保留副语言属性。我们提出使用来自同一说话人的中性情感语句对进行监督微调，并结合对声学 token 隐状态进行帧级情感蒸馏。所有改动均仅限于微调阶段，该阶段在 4 块 GPU 上耗时不到 2 小时，且不会增加推理延迟开销，同时保持具有竞争力的 180 毫秒流式延迟。在 VoicePrivacy 2024 协议下，我们的方法达到 49.2% 的 UAR（情感保留）和 5.77% 的 WER（可懂度），UAR 相较基线（39.7%→49.2%）提升了 +24%，相较情感提示变体（UAR 为 44.6%）提升了 +10%，同时保持了较强的隐私保护能力（EER 为 49.0%）。[1]

索引词：说话人匿名化、情感保留、流式语音处理、知识蒸馏

1. 引言

说话人匿名化（SA）旨在转换输入语音（来源），以隐藏说话人身份，同时保留语言内容和情感等副语言属性 [1, 2, 3]。实时 SA 对于注重隐私保护的应用至关重要，包括电话会议、呼叫中心、语音助手和在线心理健康咨询 [4],，这些场景对低延迟要求很高。尽管隐私和可懂度一直是说话人匿名化研究的首要关注点，情感保留对于自然交流同样重要。丢失情感线索会显著降低实际部署中的用户体验和沟通效果 [5]。

神经音频编解码器（NAC）语言模型的最新进展，使流式说话人匿名化在隐私与可懂度之间实现了具有竞争力的权衡 [6, 7],，不同于需要完整话语上下文的离线方法 [8, 9]。这些模型处理语音以交错的内容和声学标记序列表示，从而支持带有可控说话人身份的自回归（AR）生成。然而，情感保留面临两项根本性限制：(1) 音频续接训练范式会促使模型降低对源话语的情感保留；(2) 神经编解码器中的 VQ 瓶颈会丢弃承载情感信息的细粒度声学细节 [10, 11]。

为缓解这一局限，先前工作 [7]在推理时采用了多种情感多样的提示，这在一定程度上改善了情感保留，但显著降低了可懂度，并且需要情感标注提示；与常用于匿名化的多样化中性提示相比，这类提示更难获取。

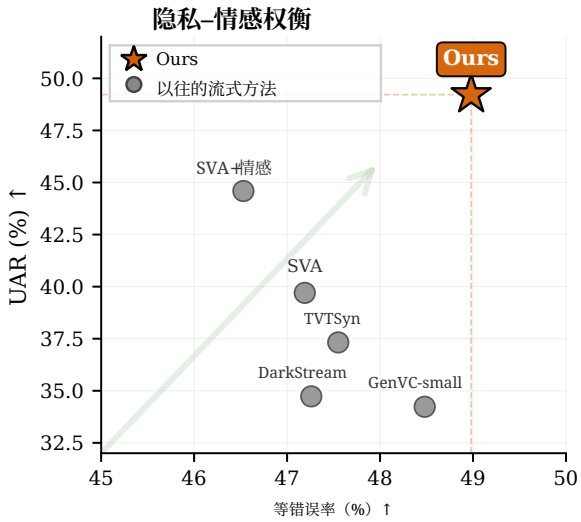


图 1：流式说话人匿名化方法在隐私与情感之间的权衡。我们的方法（橙色星形）与先前的流式方法（三角形）对比。

我们提出一种结合帧级情感蒸馏的监督微调（SFT）方法（图 2），以解决情绪退化的根本原因。我们的贡献如下：

- 我们表明，基于 NAC 的流式说话人匿名化中的情绪退化主要是训练范式问题，而非模型容量问题：重构训练对带来的提升 3×大于单独增加情感数据所带来的提升（消融实验：Exp1 与 Exp2）。
- 我们对声学 token 隐状态进行帧级情感蒸馏，将情感学习与内容监督（下一 token 预测）隔离，以避免梯度竞争。消融实验证实了这一设计选择：在非加权平均召回率和词错误率方面，声学分支蒸馏均优于语义分支（实验6 对比 Exp7）。
- 我们的方法在流式 SA 方法中实现了已报告的最高情感保留水平（UAR 为 49.2%，比基线相对提高 +24%），且**推理开销为零**，隐私保护也得到改进（图 1）。

[1]演示和代码：https://anonymous3842031239. github.io

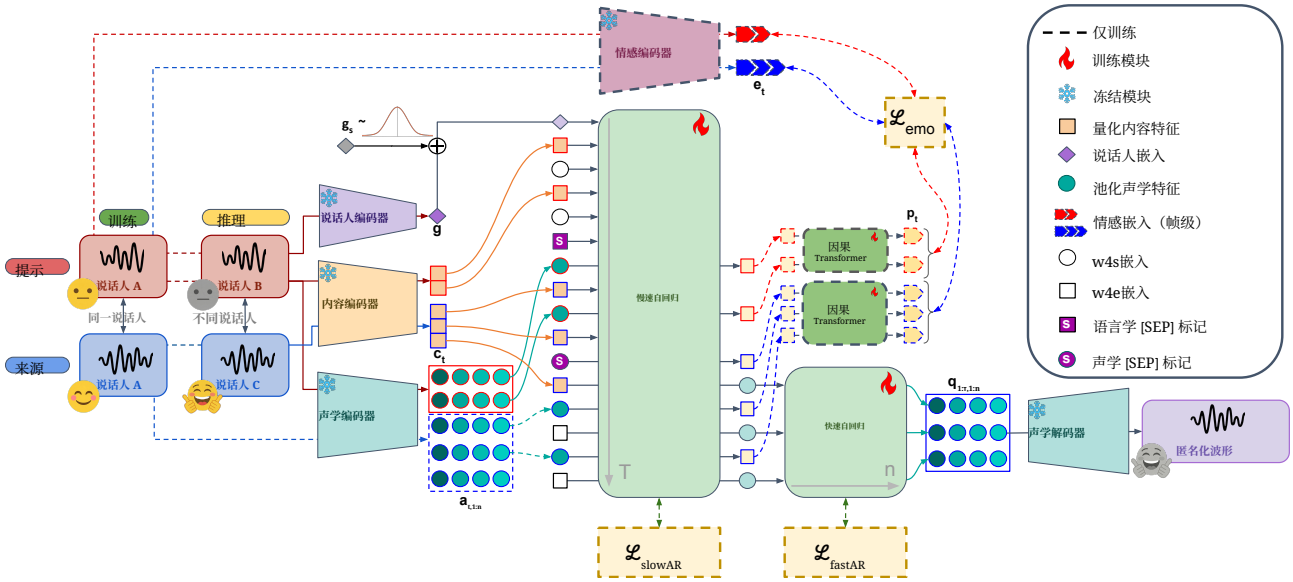


图 2： 训练与推理配置。训练：提示与来源使用同一说话人，但情感不同，迫使模型根据来源内容生成情感输出，而不是照搬提示特有的模式。在慢速自回归声学隐藏状态上进行帧级情感蒸馏（$\mathcal{L}_{emo}$，虚线），可提供额外的学习信号。推理：目标匿名说话人的中性话语掩盖来源身份，同时微调后的模型保留源情绪；与基线相比，不会增加延迟。为便于视觉呈现，图中将 *Transformer* f_θ 绘制了两次；其权重共享。

2. 相关工作

说话人匿名化。 传统的说话人匿名化方法会修改基频和共振峰等语音特征 [12, 2]，或替换说话人嵌入 [13, 14, 15, 16]。近期的神经网络方法采用语音转换 [17, 8] 或基于 NAC 的语言模型 [6, 7]。大多数方法优先考虑隐私与可懂度之间的权衡，对情感保留的关注较少 [11]。

流式说话人匿名化与语音转换。 实时说话人匿名化和语音转换系统面临延迟与质量之间的根本性权衡。StreamVoice [18] 引入了基于 NAC 的语言模型，用于交错语义与声学令牌生成式流式语音转换，并通过自回归处理实现低延迟。近期研究 [19] 指出，一个关键局限是静态说话人嵌入与动态内容表征之间存在时间失配。内容逐帧变化，而传统系统将说话人身份作为全局向量注入，导致音色过度平滑、表现力下降。时变音色（TVT）方法通过基于注意力的记忆机制，使说话人特征与内容同步变化，从而解决这一问题。GenVC [20]等自监督方法通过自回归生成，摆脱了对外部说话人编码器的依赖。这些架构层面的见解与我们的监督微调方法相辅相成；后者从训练范式的角度解决情绪退化问题。

说话人匿名化中的情感。说话人匿名化系统往往会损害情感内容 [21, 22, 23], ，其主要原因被认为是中间表示中的信息丢失 [11]。一些方法通过情感嵌入 [24] 或韵律感知匿名化 [25] 来解决这一问题。EASY [8] 通过因子化蒸馏实现了具备情感感知能力的离线说话人匿名化，在情感保留方面表现出色（非加权平均召回率为 63.8%），但需要完整的话语上下文。对于流式说话人匿名化，StreamVoiceAnon [7] 使用中性提示时，情感保留效果有限（非加权平均召回率为 39.7%），此前的工作 [26] 实现了低延迟，但隐私保护能力较弱（等错误率为 31.4%）；DarkStream [27] 的延迟为 200 毫秒，但情感保留效果有限（非加权平均召回率为 34.7%）。近期的流式语音转换研究 [19] 表明，架构改进能够提升自然度，但这些系统仍面临情感方面的挑战：内容瓶颈和说话人替换给副语言信息保留带来了额外压力。

语音知识蒸馏。 知识蒸馏已广泛应用于语音处理中，将大型预训练模型的表征迁移到更小的任务专用模型中 [28]。自监督模型的隐藏表征中包含与情感相关的信息 [29]；Emotion2Vec+ [30] 通过在大规模语音数据上进行预训练，提供帧级情感表征。与话语级情感标签不同，这些帧级特征能够捕捉每个时间步的细粒度情感动态，因此适用于流式应用。我们利用此类特征作为蒸馏目标，并仔细选择声学分支，以避免干扰语义分支上的内容监督。

3. 所提方法

3.1. 问题陈述

流式匿名化模型通常难以保留情感，这是因为存在信息瓶颈，且训练目标优先考虑内容和说话人转换，而非副语言属性。具体而言，在基于 NAC 的方法中，我们发现两个关键问题：（1）音频续接训练范式会促使模型降低对源话语情感的保留能力；（2）离散令牌表征会丢失细粒度的情感韵律。

逐情感分析显示，基线存在严重的不平衡：某些情感所占比例过高，而另一些几乎缺失（详情见第 4.3 节）。这种模式

即使使用中性提示，这种现象依然存在，表明模型在预训练期间形成了对主导声学模式的固有偏见，这很可能是由训练数据中情感分布不均衡所致。模型并未保留来源情感，而是默认采用主导声学风格。这种习得的偏见与 VQ 信息损失相结合，可能会导致系统性的情绪退化，而仅在情感数据上进行微调无法解决这一问题。

3.2. 中性-情绪配对监督微调

我们从情感语音语料库中构建训练对，其中每对都包含来自同一说话人的一条中性话语和一条情绪话语。这一设计确保：

- 说话人嵌入混合方式保持不变，因此匿名化仍然有效
- 模型必须根据中性提示生成情感声学词元
- 情感信息必须来自源内容特征，而非提示声学特征

我们还加入中性到中性的配对，以保持平衡，并防止模型认为所有源话语都带有情感。

分别为语义分支和声学分支学习的分隔词元，在每条词元流中标记提示与源序列之间的边界。这一显式边界帮助模型从提示复现过渡到源内容保留，防止其将提示特征复制到源话语中。这些词元在图 2 中显示为语言[SEP] 和声学 [SEP] 。

3.3. 帧级情感蒸馏

我们将预训练情感提取器中的帧级情绪特征蒸馏到模型的隐状态中。一个重要的设计选择是在何处应用蒸馏：

- **语义分支**：已通过下一词元预测（语言模型损失）受到监督；加入情感会造成梯度竞争 [31]
- **声学分支**：没有现有监督；可为情感学习提供干净的梯度流

我们的架构（图 2）将共享的因果 Transformer f_θ 应用于来自慢速自回归分支的声学隐藏状态 h_t^{acou}，生成预测情感嵌入 p_t，并由 $\mathcal{L}_{\text{emo}}$ 对照情绪编码器目标 e_t进行监督：

$$\mathcal{L}_{\text{emo}} = \frac{1}{T}\sum_{t=1}^{T}\|p_t - e_t\|^2, \quad p_t = f_\theta(h_t^{\text{acou}}) \tag{1}$$

其中， h_t^{acou}为声学隐藏状态， e_t 为情感编码器输出的帧级情感表示， p_t 为预测情感嵌入。虽然输出阶段仍保留 VQ瓶颈，但蒸馏损失会促使量化前的隐状态包含更丰富的情感编码。总损失由语言建模损失 [7] 和蒸馏损失组成：

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + w \cdot \mathcal{L}_{\text{emo}} \tag{2}$$

其中， $\mathcal{L}_{\text{LM}} = \mathcal{L}_{\text{slowAR}} + \mathcal{L}_{\text{fastAR}}$ 分别表示慢速自回归和快速自回归分支上的下一词元预测损失， w 控制蒸馏强度。慢速自回归沿时间轴每个时间步生成一个词元，而快速自回归使用共享权重，每个时间步自回归地生成 n 个码本词元（$q_1 \dots q_n$）。推理时， f_θ 和情感编码器会被移除；模型采用与基线相同的架构和延迟。

表1： 与先前方法的比较。 ↑/↓：数值越高/越低越好。延迟：算法延迟，单位为毫秒。*EER-L/EER-S*：懒知情/半知情攻击者。半流式：半流式处理。粗体：流式结果最佳；下划线：次佳。根据 *VoicePrivacy* 评估标准，仅纳入隐私保护效果可行的方法（*EER-L* $\geq$ *40%*）。我们的两个变体分别对应消融研究（表*2*）中的 *Exp4*（池化蒸馏）和 *Exp7*（帧级蒸馏）。

方法	Type	Lat.	WER↓	UAR↑	EER-L↑	EER-S↑
原始	–	–	1.83	70.07	5.16	–
EASY [8]	离线	–	2.70	63.81	–	45.89
GenVC-small [20]	Semi	–	8.20	34.23	48.48	15.94
DarkStream [27]	在线	200	8.75	34.73	47.26	**21.83**
TVTSyn [19]	在线	**80**	5.35	37.32	47.55	14.57
StreamVoiceAnon [7]						
vctk-1fix	在线	180	**4.54**	39.72	47.19	15.92
crema-emo-4rnd	在线	180	6.59	44.59	46.53	18.63
Ours						
池化蒸馏	在线	180	5.08	46.30	48.62	18.32
帧级蒸馏	在线	180	5.77	**49.22**	**48.98**	18.30

4. 实验

4.1. 设置

我们遵循 VoicePrivacy 2024 挑战赛协议 [3]:

- **隐私（等错误率）**： ECAPA-TDNN [32]的等错误率；包括惰性知情（在原始语音上训练）和半知情（在匿名化语音上训练）的攻击者。等错误率 = 越高，隐私越好。
- **可懂度（词错误率）**： 使用 LibriSpeech 训练的自动语音识别模型所计算的词错误率。 =越低越好。
- **情感（非加权平均召回率）**： 使用 IEMOCAP 训练的语音情感识别（SER）模型的非加权平均召回率 [33]（4 个类别：愤怒、快乐、中性、悲伤）。 = 越高越好。

我们使用 4× NVIDIA RTX 4090 GPU，以 1×10^{-4} 的学习率，在 CREMA-D [34]中性-情感配对数据上对预训练的开源 StreamVoiceAnon [7] 模型（vctk-1fix 匿名化策略，即单一固定目标说话人）进行微调，共训练 5 个 epoch。仅对 Slow AR 和 Fast AR 模块以及蒸馏变换器 f_θ （从头初始化）进行微调；所有其他组件均保持冻结。 [SEP] 标记是随机初始化的可学习嵌入。我们将 CREMA-D 筛选为四种情感（愤怒、开心、中性、悲伤），排除恐惧和厌恶类别，因为 IEMOCAP 评估集中不包含这两种情感。筛选后（基于投票分数的质量阈值为 $q = 0.5$ ），数据集包含来自 91 名演员的 7,442 个片段；我们通过多对多匹配构建了约 25,000 对中性-情感配对（将每位说话人的所有中性话语与同一说话人的所有情感及中性话语配对）。我们使用 Emotion2Vec+ large [30] （最后一层隐表示）作为蒸馏所用的情感特征提取器，权重为 $w = 0.01$；更高的权重会降低可懂度，而更低的权重带来的情感提升则逐渐减小；训练损失在第 5 至第 10 个 epoch 之间趋于平稳。需要注意的是，VoicePrivacy 2024 语音情绪识别评估器是一个基于 wav2vec 2.0 的模型 [29] ，通过 SpeechBrain 在 IEMOCAP 上训练；而蒸馏教师模型（ Emotion2Vec+）使用 data2vec 2.0 骨干网络，从而降低了循环评估的风险。该评估器属于标准 VoicePrivacy 协议的一部分，可直接与先前工作进行比较。因果变换器 f_θ 由 2 层组成，其隐藏维度与声学隐藏状态相匹配。情感评估使用 IEMOCAP 开发集和测试集。

表2： 消融研究。✓ 表示启用的组件。*StatPool/Causal*：聚合方法。蒸馏：蒸馏目标分支（*Sem* = 语义，*Aco* = 声学；– = 默认采用声学）。粗体：实验*1–7*中的最佳结果；下划线：次佳结果。所有指标均在 *IEMOCAP* 上按照 *VoicePrivacy 2024*年协议计算。

	组件						内容 ↓	情感 ↑					隐私 ↑	
模型	FT-CREMA	Neu-Emo	[SEP]	统计池化	因果	蒸馏	WER	平均	愤怒	快乐	中性	悲伤	等错误率-L	等错误率-S
基线						–	4.54	39.7	35.8	81.9	33.1	8.0	47.19	15.92
Exp1	✓					–	**5.00**	41.1	36.3	**79.6**	35.5	13.2	45.70	14.88
Exp2	✓	✓				–	5.16	45.3	35.3	75.9	48.2	21.7	47.31	16.73
Exp3	✓	✓	✓			–	5.25	47.4	34.8	72.9	50.7	31.2	47.46	16.53
Exp4	✓	✓	✓	✓		–	5.08	46.3	40.9	75.1	44.2	25.0	48.62	**18.32**
Exp5	✓	✓	✓		✓	–	5.32	48.5	40.3	65.2	**53.6**	34.8	48.19	16.78
Exp6	✓	✓	✓		✓	Sem	6.23	48.2	**48.7**	66.7	49.7	27.7	47.93	17.10
Exp7	✓	✓	✓		✓	Aco	5.77	**49.2**	38.8	62.8	52.7	**42.6**	**48.98**	18.30

4.2. 主要结果

表1展示了结果。我们方法的主要成就是情感保留：49.2% 的非加权平均召回率是所有流式方法中的最高值（SER 模型在原始语音上的得分为 70.07%，构成性能上限），同时保持了有竞争力的可懂度（5.77% 词错误率）和较强的隐私保护能力（48.98% EER-lazy）。

与采用中性提示的 StreamVoiceAnon 基线（39.7% 非加权平均召回率）相比，我们的方法在仅小幅增加词错误率的情况下（4.54% 5 77 47 19 → 48 98 . %），实现了 +24% 的相对非加权平均召回率提升，并改善了隐私性（ EER-lazy: . % . %）。我们推测，情感蒸馏有助于在声学隐藏状态中解耦情感信息与说话人信息，从而减少身份泄露。与情感提示变体（SVA+EMO，44.6% 非加权平均召回率）相比，我们实现了 +10% 的相对非加权平均召回率，同时具有更优的词错误率（6.59% → 5.77%）和隐私性。值得注意的是，我们仅使用单个中性提示，而非四种多种情感提示。

与其他隐私保护效果可行的流式方法（EER-lazy >40%）相比，我们的方法相较于 GenVC-small（34.2%）、DarkStream（34.7%）和 TVTSyn（37.3%），将非加权平均召回率相对提升了 +32–44%，且词错误率和隐私保护效果相当或更佳，表明训练范式方法可与架构改进相辅相成。

4.3. 消融研究

表2系统地评估了各个组件。从 Exp1 开始的所有实验均在 CREMA-D 上进行微调；从 Exp2 开始加入中性-情感配对；从 Exp3 开始加入分隔标记；Exp4–7 则探索不同的蒸馏架构。

在情感数据上微调（Exp1）：仅在 CREMA-D 上进行微调，非加权平均召回率仅为 +1.4，证实问题并非出在数据域，而是训练范式。

中性-情感配对（Exp2）：促使模型根据中性提示生成情感输出，带来了最大幅度的单项提升（非加权平均召回率提高 +4.1），部分验证了我们关于提示驱动的情感退化的假设。悲伤情感的识别率从 8.0% 提高到 21.7%。

分隔标记（Exp3）：在提示与源序列之间进行显式边界标记，使非加权平均召回率提高 +2.1，并进一步将悲伤提升至31.2%。

情感蒸馏架构（实验4-5）：我们探索了蒸馏头的两种聚合策略。统计池化（实验4）在预测前对隐状态计算均值和标准差，取得了 46.3% 的非加权平均召回率和 5.08% 的词错误率。然而，实验4的非加权平均召回率实际上低于实验3的 47.4%，这表明话语级池化会丢失情绪时序动态。因果变换器（实验5）保留了帧级预测，取得了 48.5% 的非加权平均召回率和 5.32% 的词错误率，证实时序建模能更好地捕捉帧级情感动态。

蒸馏目标选择（实验6 vs Exp7）：蒸馏至语义隐藏状态（实验6）达到48.2%的非加权平均召回率，但由于与内容监督存在梯度竞争，词错误率恶化至6.23%。声学蒸馏（Exp7）取得更高的非加权平均召回率（49.2%），同时词错误率更优（5.77%），证实声学分支提供了更干净的梯度流。

逐情感分析：最显著的提升出现在“悲伤”上：从8.0%（基线）提升至42.6%（我们的方法）。中性情绪从33.1%提升至52.7%（+59%）。“高兴”的下降（81.9% → 62.8%）代表对过度预测偏差的校正，因为基线无论源情绪如何，都默认输出听起来高兴的结果。

5. 结论

我们提出了一种用于情感保留流式说话人匿名化的 SFT 方法。消融实验表明，对于该架构而言，情感保留在很大程度上是一个训练范式问题：重构训练对（Exp1→Exp2, +4 2 3× . UAR）带来的提升，大于仅添加情感数据（Baseline→Exp1, +1.4 UAR），这说明模型具备编码情感的能力，只是缺少恰当的训练信号。结合对声学隐状态进行帧级情感蒸馏后，我们的方法将基线的非加权平均召回率从 39.7% 提升至 49.2%（相对提升 +24%），同时也改善了隐私保护（EER-lazy：47.2% → 49.0%），且不会增加任何推理开销。

与离线方法（EASY：63.8% 非加权平均召回率）相比，仍存在差距，这反映了延迟与质量之间根本性的权衡：离线方法可以访问完整的话语上下文，从而支持双向建模和话语级情绪优化。未来工作将探索在因果约束下利用更长范围的情感上下文、扩展至维度情绪模型，以及针对自适应攻击进行对抗训练。尽管我们遵循标准的 VoicePrivacy 2024 评估协议，但本研究仍有局限，包括仅依赖单一语音情绪识别评估器、缺少主观听音测试，以及仅在表演式语音语料库上进行评估（在 CREMA-D 上微调，并在 IEMOCAP 上评估）。未来工作应在自然产生的情感语料库（如 MSP-Podcast）上进行验证。

6. 生成式人工智能使用声明

我们使用了生成式人工智能工具辅助稿件编辑和格式编排。所有技术内容、实验设计和分析均由作者完成。

7. 参考文献

[1] N. Tomashenko、X. Wang、E. Vincent、J. Patino、B. M. L. Srivastava、P.-G. No´e、A. Nautsch、N. Evans、J. Yamagishi、B. O'Brien、A. Chanclu、J.-F. Bonastre、M. Todisco 和 M. Maouche，"VoicePrivacy 2020 挑战赛：结果与发现"，*Computer Speech& Language*，第 74 卷，第 101362 页，2022年。[2] M. Panariello、N. Tomashenko、X. Wang、X. Miao、P. Champion、H. Nourtel、M. Todisco、N. Evans、E. Vincent 和 J. Yamagishi，"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 3497–3512 页，2024年。[3] N. Tomashenko、X. Miao、P. Champion、S. Meyer、M. Panariello、X. Wang、N. Evans、E. Vincent、J. Yamagishi 和 M. Todisco，"第三届 VoicePrivacy 挑战赛：在语音匿名化中保留情感表达能力和语言内容"，*arXiv* 预印本 *arXiv:2601.11846*，2026年。[4] S. Meyer 和 N. T. Vu，"语音匿名化的应用场景"，载于第 *5* 届语音通信安全与隐私研讨会（*SPSC*）论文集，2025年。

[5] Z. Cai, H. X. Li, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, M. Andrews 和 M. Wiesner，《情感保持说话人匿名化中的隐私与情感保持权衡》，发表于 *IEEE*口语技术研讨会（*SLT*）论文集，2024年，第 409–414 页。

[6] M. Panariello、F. Nespoli、M. Todisco 和 N. Evans，"使用神经音频编解码器语言模型的说话人匿名化"，载于*ICASSP 2024*年论文集，2024年，第4725–4729页。[7] N. Kuzmin、S. Liu、K. A. Lee 和 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*ICASSP 2026*年论文集，2026年。[8] J. Yao、H. Liu、E. S. Chng 和 L. Xie，"EASY：通过因子化蒸馏实现情感感知说话人匿名化"，载于*Interspeech 2025*年论文集，2025年，第3219–3223页。[9] J. Yao、N. Kuzmin、Q. Wang、P. Guo、Z. Ning、D. Guo、K. A. Lee、E.-S. Chng 和 L. Xie，"NPU-NTU 2024年语音隐私挑战赛系统"，载于第四届语音通信安全与隐私研讨会，2024年，第67–71页。[10] X. Zhang、D. Zhang、S. Li、Y. Zhou 和 X. Qiu，"SpeechTokenizer：面向语音语言模型的统一语音分词器"，载于国际学习表征会议（*ICLR*）论文集，2024年。[11] ¨U. E. Gaznepoglu 和 N. Peters，"为什么基于解耦的说话人匿名化系统无法保留情感？"，载于*ICASSP 2025*年论文集，2025年，第1–5页。[12] J. Patino、N. Tomashenko、M. Todisco、A. Nautsch 和 N. Evans，"使用 McAdams系数的说话人匿名化"，载于*Interspeech 2021*年论文集，2021年，第1099–1103页。[13] B. M. L. Srivastava、N. Tomashenko、X. Wang、E. Vincent、J. Yamagishi、M. Todisco、A. Nautsch、J. Shan 和 N. Evans，"基于x-vector的说话人匿名化设计选择"，载于*Interspeech 2020*年论文集，2020年，第2512–2516页。[14] B. M. L. Srivastava、M. Maouche、M. Sahidullah、E. Vincent、A. Bellet、M. Tommasi、N. Tomashenko、X. Wang 和 J. Yamagishi，"基于x-vector的说话人匿名化的隐私与效用"，*IEEE/ACM* 音频、语音与语言处理汇刊，第30卷，第2383–2395页，2022年。[15] N. Kuzmin、H.-T. Luong、J. Yao、L. Xie、K. A. Lee 和 E.-S. Chng，"NTU-NPU 2024年语音隐私挑战赛系统"，载于第四届语音通信安全与隐私研讨会（*SPSC*）论文集，2024年，第72–79页。[16] S. Meyer、P. Tilli、P. Denisov、F. Lux、J. Koch 和 N. T. Vu，"使用生成对抗网络进行语音匿名化以保护说话人隐私"，载于*IEEE* 口语语言技术研讨会（*SLT*）论文集，2023年，第912–919页。[17] S. Meyer、F. Lux、J. Koch、P. Denisov、P. Tilli 和 N. T. Vu，"韵律不是身份：一种使用韵律克隆的说话人匿名化方法"，载于*ICASSP 2023*年论文集，2023年，第1–5页。[18] Z. Wang、Y. Chen、X. Wang、L. Xie 和 Y. Wang，"StreamVoice：面向实时零样本语音转换的可流式上下文感知语言建模"，载于计算语言学协会第*62*届年会论文集（第一卷：长篇论文）。泰国曼谷：计算语言学协会，2024年，第7328–7338页。[19] 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"，载于国际学习表征会议（*ICLR*）论文集，2026年。[20] Z. Cai、H. X. Li、A. Garg、L. P. García-Perera、K. Duh、S. Khudanpur、M. Wiesner 和 N. Andrews，"GenVC：自监督零样本语音转换"，载于*IEEE* 自动语音识别与理解研讨会（*ASRU*）论文集。IEEE，2025年。[21] S. Ghosh、A. Das、Y. Sinha、I. Siegert、T. Polzehl 和 S. Stober，"Emo-StarGAN：半监督任意到多目标的非平行情感保持语音转换"，载于*Interspeech 2023*年论文集，2023年，第2093–2097页。[22] A. Das、S. Ghosh、T. Polzehl、I. Siegert 和 S. Stober，"StarGAN-VC++：使用深度嵌入实现情感保持语音转换"，载于第十二届*ISCA*语音合成研讨会（*SSW*）论文集，2023年，第81–87页。[23] H. Hua、Z. Shang、X. Li、P. Shi、C. Yang、L. Wang 和 P. Zhang，"情感语音匿名化：在伪说话人语音生成中保留情感特征"，载于第四届语音通信安全与隐私研讨会（*SPSC*）论文集，2024年，第55–60页。[24] X. Miao、Y. Zhang、X. Wang、N. Tomashenko、D. C. L. Soh 和 I. Mcloughlin，"调整通用基于解耦的说话人匿名化以增强情感保留"，*Computer Speech&Language*，第94卷，文章101810，2025年。[25] J. He、S. Zhao、J. Zhou、H. Sun、H. Wang 和 Y. Qin，"用于语音隐私保护的情感保持韵律匿名化网络"，载于*ICASSP 2025*年论文集，2025年，第1–5页。[26] 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*）论文集。IEEE，2024年。[27] 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。[28] H.-J. Chang、S.-w. Yang 和 H.-y. Lee，"DistilHuBERT：通过隐单元BERT的逐层蒸馏进行语音表征学习"，载于*ICASSP 2022*年论文集，2022年，第7087–7091页。[29] L. Pepino、P. Riera 和 L. Ferrer，"使用wav2vec 2.0嵌入进行语音情感识别"，载于*Interspeech 2021*年论文集，2021年，第3400–3404页。[30] Z. Ma、Z. Zheng、J. Ye、J. Li、Z. Gao、S. Zhang 和 X. Chen，"emotion2vec：面向语音情感表征的自监督预训练"，载于计算语言学协会论文集：*ACL 2024*年。泰国曼谷：计算语言学协会，2024年，第15747–15760页。[31] Y. Hu、C. Chen、R. Li、Q. Zhu 和 E. S. Chng，"端到端抗噪语音识别中的多任务学习梯度补救"，载于*ICASSP 2023*年论文集，2023年，第1–5页。[32] B. Desplanques、J. Thienpondt 和 K. Demuynck，"ECAPA-TDNN：基于时延神经网络的说话人验证中的增强通道注意力、传播与聚合"，载于*Interspeech 2020*年论文集。

, 2020, 第3830–3834页。

[33] C. Busso、M. Bulut、C.-C. Lee、A. Kazemzadeh、E. Mower、S. Kim、J. N. Chang、S. Lee 和 S. S. Narayanan，“IEMOCAP：交互式情感二人动作捕捉数据库”，语言资源与评测，第42卷，第4期，第335–359页，2008年。[34] H. Cao、D. G. Cooper、M. K. Keutmann、R. C. Gur、A. Nenkova 和 R. Verma，“CREMA-D：众包情感多模态演员数据集”，*IEEE*情感计算汇刊，第5卷，第4期，第377–390页，2014年。