第三届 VoicePrivacy 挑战赛：在语音匿名化中保留情感表现力和语言内容

Natalia Tomashenko[a]、* Xiaoxiao Miao[b]、Pierre Champion[c]、Sarina Meyer[d]、Michele Panariello[e]、Xin Wang[f]、Nicholas Evans[e]、Emmanuel Vincent[a]、Junichi Yamagishi[f] 和 Massimiliano Todisco[e]

[a]洛林大学、法国国家科学研究中心、法国国家信息与自动化研究所、洛林计算机科学与应用研究实验室，*F-54000*，南锡，法国[b]昆山杜克大学，昆山，*215316*，中国[c]法国国家信息与自动化研究所国防与安全部，格勒诺布尔，法国[d]斯图加特大学自然语言处理研究所，德国[e]音频安全与隐私组，*EURECOM*，比奥，索菲亚科技园，*F-06904*，法国[f]日本国立信息学研究所，千代田区，东京，*101-8340*，日本

arXiv:2601.11846v1 [cs.CL] 17 Jan 2026

文章信息

关键词：隐私 匿名化 攻击模型 VoicePrivacy 挑战赛 说话人验证 语音识别 情感识别 语音转换 效用 评估指标

摘要

本文介绍了2024年举办的第三届 VoicePrivacy 挑战赛的结果与分析，该挑战赛旨在推动语音匿名化技术的发展。任务是开发一种用于语音数据的语音匿名化系统，在隐藏说话人声音身份的同时保留语言内容和情绪状态。我们系统地概述了挑战赛框架，详细介绍了匿名化任务，以及用于系统开发和评估的数据集。我们阐述了攻击模型，以及用于评估隐私保护（隐藏说话人语音身份）和效用（保留内容和情绪状态）的客观评估指标。我们介绍了六种基线匿名化系统，并总结挑战赛参赛者开发的创新方法。最后，我们提出关键见解与观察结果，以指导未来 VoicePrivacy 挑战赛的设计，并指出语音匿名化研究中值得探索的方向。

1. 引言

语音数据属于隐私法规的管辖范围，例如欧盟《通用数据保护条例》（GDPR）。这是因为语音录音包含大量个人和敏感信息，例如说话人的身份、年龄和性别、健康状况、人格特征、种族或民族出身、地理背景、社会身份以及社会经济地位（Nautsch 等人，2019年）。VoicePrivacy倡议及其挑战赛系列（Tomashenko 等人，2020b）于2020年发起，旨在引领社区共同开发语音技术的隐私保护解决方案。

隐私保护需要结合多种互补方案，不仅要掩盖说话人的身份，还要掩盖特定的语言内容、语言外特征和背景声音，因为这些信息也可能泄露说话人的身份或其他敏感信息。迄今为止，VoicePrivacy语音匿名化方面的工作重点是开发相关方案，在隐藏声音身份的同时，保留与一个或多个下游任务（例如自动语音识别）相关的一系列其他属性。研究人员已确定并设计了通用数据集和实验方案，以便使用一组评估指标，对各种匿名化方案进行公平的基准测试。VoicePrivacy 前两届分别于 2020 年（Tomashenko 等人，2020b，2022c，

*通讯作者：natalia.tomashenko@inria.fr (N. Tomashenko)；xiaoxiao.miao@dukekunshan.edu.cn (X. Miao)；pierre.champion@inria.fr (P. Champion)；sarina.meyer@ims.uni-stuttgart.de (S. Meyer)；michele.panariello@eurecom.fr (M. Panariello)；wangxin@nii.ac.jp (X. Wang)；evans@eurecom.fr (N. Evans)；emmanuel.vincent@inria.fr (E. Vincent)；jyamagis@nii.ac.jp (J. Yamagishi)；massimiliano.todisco@eurecom.fr (M. Todisco)https://sites.google.com/view/natalia-tomashenko (N. Tomashenko)；https://xiaoxiaomiao323.github.io/(X. Miao)；https://www.eurecom.fr/en/people/evans-nicholas (N. Evans)；https://members.loria.fr/EVincent/ (E. Vincent)；https://yamagishilab.jp/ (J. Yamagishi)；https://www.eurecom.fr/en/people/todisco-massimiliano (M. Todisco) 开放研究者与贡献者身份识别码：0000-0002-7125-2382 (N. Tomashenko)；0000-0002-6645-6524 (X. Miao)；0000-0003-1225-7957 (P. Champion)；0009-0004-1117-4783 (S. Meyer)；0009-0007-4154-5460 (M. Panariello)；0000-0001-8246-0606 (X. Wang)；0000-0002-8459-1041 (N. Evans)；0000-0002-0183-7289 (E. Vincent)；0000-0003-2752-3955 (J. Yamagishi)；0000-0003-2883-0324 (M. Todisco)

2021，2020a，2022b）和 2022 年（Tomashenko 等人，2022a，b；Panariello 等人，2024b），推动了语音匿名化领域的显著进展，并促成了一个新的研究社群的形成。

本文介绍了最近一届于 2024年 举办并落下帷幕的 VoicePrivacy 挑战赛（VPC）。与往届一样，2024年 这一届继续聚焦语音匿名化，以及语言内容和副语言属性的保留。2024年 这一届新增了一项具体要求：保留说话人的情绪状态。这是语音匿名化在许多现实应用场景中的一项关键副语言属性，例如在呼叫中心使用第三方语音分析服务。[1]

强调在语音匿名化中保留情感属性，增加了这项任务本身的难度；它要求在隐藏声音身份的同时，保留日益丰富的语言和副语言特征。因此，后续挑战赛届次应致力于拓宽乃至转变其范围，不仅要更明确地关注匿名化后必须抑制或掩盖哪些内容，也要同样重视应保留哪些内容。引入这类平衡且逐步提高难度的约束，有助于推动科学进步，促进语音表征解耦，并推动形成更稳健、更有原则的语音匿名化方法。新的关注重点还要求采用更多情绪标注数据库和实验方案，并制定新的指标，以支持对情感保留效果的评估。

三个 VoicePrivacy 挑战赛届次之间的差异总结于表 1中。表中突出展示了 2020年 至 2024年间，目标、攻击者能力、评估策略和方法多样性的逐步演进：VPC 2020 的重点是在保留语言内容、自然度和语音独特性（作为补充目标）的同时，隐藏说话人语音身份（针对多种攻击场景： 无知、有限知情、半知情）。VPC 2022 加强了对 半知情攻击者的防御，并引入了语调保持；VPC 2024 则将要求扩展至同时保留语言内容和情感表现力。这一演进体现了攻击者模型日益先进，评估指标的范围不断扩大——从基础效用度量扩展至韵律和情感保留——以及匿名化方法的多样性显著提升：提交的系统从在基线上的渐进改进，发展到采用神经编解码器、先进的属性解耦方法和混合方法，以灵活平衡隐私-效用权衡。这些进展凸显了该领域对稳健、多维语音隐私解决方案日益提高的要求。

本文介绍 VPC 2024，阐述挑战赛任务、攻击模型、协议、匿名化基线和评估指标。此外，我们总结了挑战赛参与者设计的36种匿名化系统，进一步扩展了在 VoicePrivacy 2024研讨会（与第四届语音通信安全与隐私研讨会（SPSC）同期举行）上所呈现的分析；[2] 该联合活动与 Interspeech 2024 在同一地点举行。我们还将这些系统与历届 VPC 中涌现的系统进行新的比较，并进一步探讨在更强攻击面前隐私评估的局限性及其可靠性。最后，我们概述挑战赛后结果，并分享对未来研究的看法，以及暂定于2026年举行的下一届 VPC 可能的发展方向。

2. 挑战设计

在本节中，我们　介绍挑战赛设置概况：匿名化任务、攻击模型和数据　　.

2.1. 语音匿名化任务

隐私保护被表述为一场博弈：一方是用户，其共享数据是为了完成某项预期的下游任务；另一方是攻击者，其访问并使用这些数据或由此衍生的数据，以推断有关数据主体的信息（Qian 等人，2018年；Srivastava 等人，2020b；Tomashenko 等人，2020b）。在此，我们考虑这样一种情境：用户共享匿名化语音话语，以用于下游自动语音识别（ASR）和语音情感识别（SER）任务；而攻击者则试图利用相同的匿名化语音话语识别说话人（见图1）。

用户共享的话语称为试验 话语。为了隐藏自己的身份，用户会在共享之前，先使用语音匿名化系统处理每条话语。处理后的话语听起来仿佛由另一位说话人说出，这位说话人称为伪说话人。例如，伪说话人可能是一个与任何真实个体的声音都不对应的人工声音。挑战赛参与者的任务是开发这一语音

[1]这项要求体现了一种选择性匿名化方法：为确保符合 GDPR，抑制说话人身份线索，同时有意保留情感等某些副语言属性，以支持下游效用。通常认为，只要保留的情感无法与个人关联，也不会泄露敏感的健康信息，这种做法就符合 GDPR。[2] https://spsc-symposium.de/2024

表1VoicePrivacy 挑战赛（2020–2024年）对比

VoicePrivacy 2020	VoicePrivacy 2022	VoicePrivacy 2024
Task		
移除说话人标识 保留语言内容 保留自然度 & 可懂度	移除说话人标识 保留语言内容 **保留语调** 保留自然度 & 可懂度	移除说话人标识 保留语言内容 **保留情绪状态**
训练数据		
LibriSpeech：train-clean-100、train-other-500 LibriTTS：train-clean-100、train-other-500 VoxCeleb-1、2	LibriSpeech：train-clean-100、train-other-500 LibriTTS：train-clean-100、train-other-500 VoxCeleb-1、2	**开放资源（数据、模型）** **（基于参与者提案）**
开发与评估数据		
LibriSpeech 开发集/测试集-clean VCTK	LibriSpeech 开发集/测试集-clean VCTK	LibriSpeech 开发集/测试集-clean **IEMOCAP**
测试数据的匿名化级别		
说话人	说话人	**话语**
客观评估		
攻击模型		
无知，有限知情，半知情	半知情	半知情
隐私度量指标		
主要指标：等错误率（EER）， 评估后：C_{llr}, C^{min} 对数似然比，可链接性， ZEBRA（Nautsch 等人，2020年）， 去身份化（DeID）（Noé等人，2020年，2022年）	EER	EER
效用指标		
词错误率（在原始数据和匿名化数据上训练的自动语音识别） 语音独特性增益（G_{VD}）	词错误率（在**匿名化** 数据上训练的自动语音识别） 语音独特性增益（G_{VD}） **音高相关性**（$\rho_{F0} > 0.3$）	词错误率（基于**原始**数据训练的自动语音识别） **非加权平均召回率（基于原始数据训练的语音情感识别）**
主观评估		
隐私度量指标		
主要指标：说话人可验证性 评估后：可链接性（O'Brien 等人，2021年）	说话人可验证性	—
效用指标		
自然度与可懂度	自然度与可懂度	—
排名策略		
—	**按隐私类别内的效用排序** **类别（等错误率 ≥ 15%、20%、25%、30%）**	按隐私类别内的效用排序 类别（**等错误率 ≥ 10%、20%、30%、40%**）
基线		
– 神经语音转换： B1（x向量 + 神经声码器） – 信号处理： B2（McAdams 固定系数）	– 神经语音转换： **B1.a、B1.b** – 信号处理： B2（McAdams **随机**系数）	B1（来自 VPC2022 的 B1.b） B2（McAdams 随机系数） **B3（语音转写 + 文本转语音和生成对抗网络）** **B4（神经编解码器）** **B5、B6（采用矢量量化的 ASR-BN）**
提交的系统		
– 神经语音转换： 属性解耦 （重点关注 x-vector 匿名化， 主要是在基线上逐步改进 优于基线） – 信号处理： 共振峰、基频和语速修改	– 神经语音转换： 属性解耦 **(修改基线中的所有组件**： 内容、说话人和韵律 特征提取；说话人和韵律 匿名化器；语音生成模型） – 信号处理： 共振峰、基频和语速修改	– 神经语音转换： 属性解耦 （说话人、语言、基频、**情感**； **多种情感保留策略** **（预训练编码器、蒸馏）；kNN-VC** –神经编解码器 –自动语音识别+文本转语音级联 –混合方法

匿名化系统应当：(a) 输出语音波形；(b) 在话语级隐藏声音身份；(c)保留语言和情感内容。

要求 (b) 所述的话语级 匿名化，意味着语音匿名化系统必须将声音身份替换为伪说话人的声音身份，且独立处理每个话语。伪说话人分配过程必须在所有话语中保持一致，且不得依赖说话人标签。如果该过程使用随机数生成器，则为每个话语生成的随机数必须不同，从而为每个话语分配不同的伪说话人。或者，也可以为所有 话语分配同一个伪说话人，以满足此项要求。要求 (c) 的实现情况通过效用 指标进行评估。具体而言，我们测量

使用基于原始（未受保护）数据训练的自动语音识别（ASR）和语音情感识别（SER）系统，分别通过词错误率（WER）和非加权平均召回率（UAR）衡量语言内容和情感内容的保留情况。

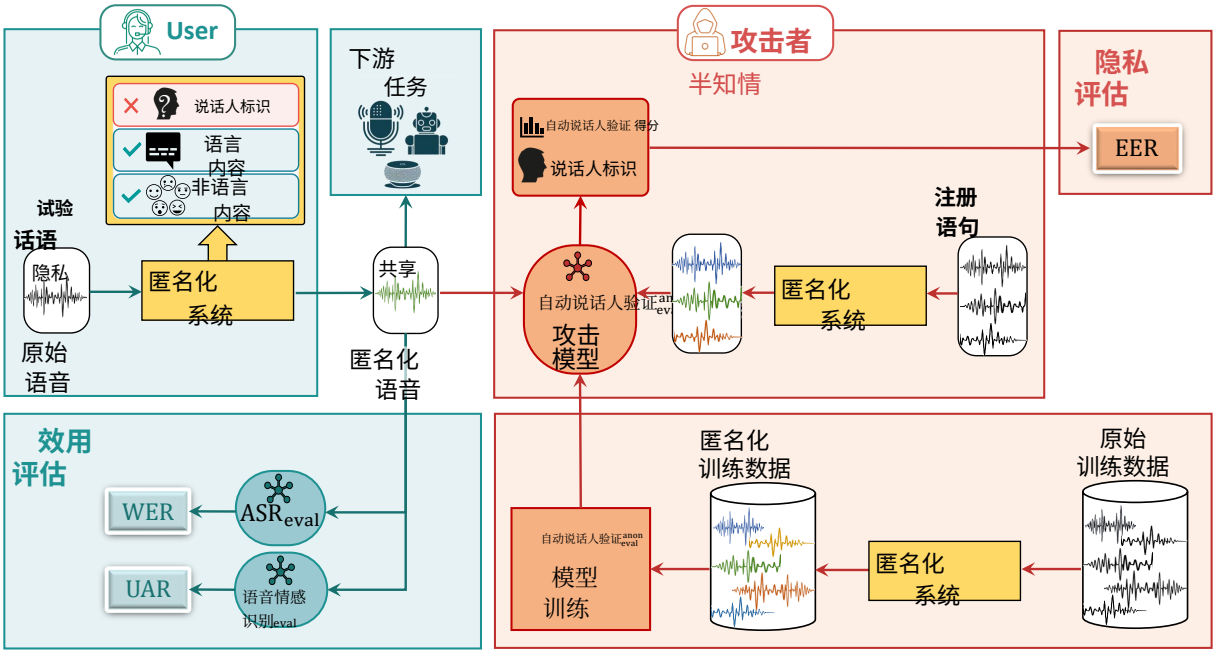


图 1： 隐私保护场景被建模为用户 与攻击者 之间的一场博弈：当说话人身份被视为需要通过匿名化加以保护的个人信息时，应保留语言和情感内容，以供下游任务使用。使用半知情 攻击模型 ASV^{anon} 评估 和等错误率（EER）指标，对匿名化数据进行隐私评估。效用评估则使用：(1) 自动语音识别（ASR）模型 ASR_{eval} 和词错误率（WER）指标；(2) 语音情感识别（SER）模型 SER_{eval} 和非加权平均召回率（UAR）指标。

2.2. 攻击模型

对于每位目标说话人，假设攻击者 能够获取该说话人所说的话语，这些话语称为注册 话语。攻击者利用这些话语部署自动说话人验证（ASV），以重新识别每条匿名化试验 话语中所包含语音对应的说话人。

我们假设攻击者可以获取：(a) 每位说话人的若干条注册语句；(b) 用户所采用的同一语音匿名化系统。利用这些资源，攻击者会对注册语句进行匿名化，以减少其与测试语句之间的不匹配，并训练一个适配该特定匿名化系统的自动说话人验证系统。与用户一样，攻击者在话语级进行语音匿名化。迄今为止，这一攻击模型在概念上是已知最强的，因此我们将其用于隐私评估。

2.3. 数据与预训练模型

语音匿名化系统的训练、开发和评估均使用公开可用的资源。开发与评估数据是固定的，而训练资源的选择则具有灵活性，但须遵守下文所述的限制。

2.3.1. 训练资源

除以往挑战赛届次所使用的训练数据（*VoxCeleb-1*、*2*；*LibriSpeech train-clean-100*、*train-other-500*；*LibriTTStrain-clean-100*、*train-other-500*）以及用于构建和开发基线匿名化系统的数据和预训练模型（语料库：*RAVDESS*、*ESD*；模型：*HuBERTBase*、*EnCodec*、*Bark*和*wav2vec2_large_west_germanic_v2*）之外——详情请参见第3.4 节——参赛者还可以提出其他训练资源，包括数据和预训练模型。根据收到的建议，

表 2用于自动说话人验证和自动语音识别评估的*LibriSpeech* 开发集和评估集统计

on.

	子集	Type	说话人数			话语数
			女性	Male	总计	
开发	LibriSpeech dev-clean	注册	15	14	29	343
		试验	20	20	40*	1,978
评估	LibriSpeech 测试集（干净语音）	注册	16	13	29	438
		试验	20	20	40*	1,496

*包含仅出现在不同说话人 试验中的11位说话人。

表3用于SER评估的*IEMOCAP* 开发集和评估集的构建与统计数据。训练集 子集指SER评估系统的训练数据。

对话		#话语数	第1折	第2折	第3折	第4折	第5折
第1场次	女性	528	Dev	训练集	训练集	训练集	训练集
	Male	557	Eval				
第2场次	女性	481	训练集	Eval			
	Male	542		Dev			
第3场次	女性	522		训练集	Dev		
	Male	629			Eval		
第4场	女性	528			训练集	Eval	
	Male	503				Dev	
第5场	女性	590				训练集	Eval
	Male	651					Dev

允许使用的训练数据和预训练模型清单已最终确定，并通过更新评估计划予以发布（Tomashenko 等人，2024a）。该清单见附录8 中的表A。

2.3.2. 开发与评估数据

开发和评估数据集来自两个语料库：*LibriSpeech*[3] (Panayotov 等人，2015年）和 *IEMOCAP*(Busso 等人，2008），如表 2 和 3所示。 *LibriSpeech* 包含 960 小时的英语朗读语音（16 kHz），以有声读物形式呈现，通常用于 ASV 和自动语音识别评估。 *LibriSpeech* 的评估集和开发集与往届挑战赛相同。 *IEMOCAP* 包含 12 小时的情感语音（16 kHz），采集自 10 名英语演员（5 名女性、5 名男性）之间的双人对话。根据 Pappagari 等人(2020年），我们选取 4 种情感的子集进行 SER 评估： 中性、悲伤、愤怒和 快乐 （其中将快乐和 兴奋 类别合并）。由于数据局限性，我们采用留一对话交叉验证，使用四段对话（8 名说话人）进行 SER 训练，剩余一段对话（2 名说话人）用于开发和评估。

3. 隐私与效用评估

提交系统的客观排名采用三项指标：作为隐私度量的等错误率（EER），以及两项效用指标：词错误率（WER）和非加权平均召回率（UAR）。这些指标依托自动说话人验证（ASV）、自动语音识别（ASR）和语音情感识别（SER）系统。

[3]LibriSpeech: http://www.openslr.org/12

表4*LibriSpeech* 训练集的统计数据，用于自动说话人验证和自动语音识别评估模型。

系统	子集	时长（小时）	说话人数			#话语
			女性	Male	总计	
ASV	LibriSpeech train-clean-360	363.6	439	482	921	104,014
ASR	LibriSpeech train-960	960.9	1128	1210	2338	281,241

表 5用于评估的同说话人和不同说话人配对数量。

子集		试验	女性	Male	总计
开发	LibriSpeech dev-clean	同说话人	704	644	1,348
		不同说话人	14,566	12,796	27,362
评估	LibriSpeech 测试集（干净语音）	同说话人	548	449	997
		不同说话人	11,196	9,457	20,653

3.1. 评估模型的训练数据

自动说话人验证系统在 *LibriSpeech-train-clean-360* 上训练，自动语音识别系统则在完整的 *LibriSpeech-train-960*数据集上训练，其统计信息见表 4。每个*IEMOCAP* 交叉验证折对应的语音情感识别系统都在相应的 *IEMOCAP* 训练子集上训练，其统计信息见表3。训练和评估均使用所提供的配方和模型。[4] 隐私评估所用的模型由参赛者使用所提供的训练脚本，在其匿名化训练数据上训练；效用评估所用的模型则由组织方提供。

3.2. 隐私度量：等错误率（等错误率）

用于隐私评估的自动说话人验证系统是一个 ECAPA-TDNN 模型（Desplanques 等人，2020年），其卷积帧层包含 512 个通道。该模型通过将 *SpeechBrain* （Ravanelli 等人，2021年）的 VoxCeleb 数据集配方调整至 *VoxCeleb* 数据集应用于 *LibriSpeech* 数据集而实现。如图 1所示，假设攻击者可以访问正在评估的匿名化系统（Srivastava 等人，2020年b；Tomashenko 等人，2022年c）。攻击者使用该系统匿名化注册语音，以减少注册语音与试验语音之间的不匹配。此外，攻击者使用同一系统匿名化 *LibriSpeech-train-clean-360* 数据集，并重新训练自动说话人验证系统（记为 ASV^{anon} eval），使其能够适配处理经过类似匿名化的输入。在所有情况下，匿名化均在话语级应用，并使用与试验语音匿名化相同的伪说话人分配过程。对于拥有多个注册语音的说话人，系统会为每段话语提取嵌入向量，然后取其平均值。

对从注册和测试语句中提取的每一对嵌入向量计算余弦相似度分数，并通过设置判决阈值生成同说话人和不同说话人判定。将阈值 θ 下的自动说话人验证虚警率和漏检率分别记为 $P_{\text{fa}}(\theta)$ 和 $P_{\text{miss}}(\theta)$，等错误率（EER）定义为使两个检测错误率相等的判决阈值 θ_{EER}，即 $\text{EER} = P_{\text{fa}}(\theta_{\text{EER}}) = P_{\text{miss}}(\theta_{\text{EER}})$。EER 越高，匿名化效果越好。用于估计 EER 的同说话人和不同说话人注册/测试语句对数量见表5，其中列出了每个数据集的开发集和评估分区。

[4]评估脚本：https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

[5] 在此语境中，隐私特指说话人去身份化（以自动说话人验证等错误率增幅衡量），而效用表示内容和情感保留（自动语音识别词错误率、情感识别非加权平均召回率）；两者共同构成匿名化性能。从更广泛的意义上说，隐私涵盖对所有可能识别或关联到个人的个人信息和敏感信息的保护，包括元数据、上下文细节，以及语音特征以外的行为模式。因此，即使说话人去身份化达到完美水平，如果说话人姓名、位置或交互模式等其他可识别信息遭到暴露或披露，也无法保证隐私得到全面保护。

3.3. 效用指标

2022年 VPC 有三个效用指标，而2024年版本只有两个。语言内容保留的度量保持不变，基频相关性和语音独特性增益的度量则被情感保留度量所取代。基频相关性的度量被认为提供的信息或可靠性不足，而语音独特性增益涉及过于狭窄的使用场景。说话人的情绪状态是一项关键的副语言属性，与更广泛的应用场景相关。

3.3.1. 词错误率（***WER***）

使用自动语音识别系统评估语言内容的保留情况[6]（记作 ASR_{eval}），该系统使用*LibriSpeech-train-960* 数据集和 *wav2vec2-large-960h-lv60-self*[7] 表征进行微调，并采用 *SpeechBrain* 配方。自动语音识别评估模型使用原始（未经处理的）数据进行训练。然而，与先前的 VPC 版本不同，该模型随后保持固定，而不针对匿名化数据进行调整。

使用原始数据训练自动语音识别系统以估算效用，旨在最大限度地减少声学失真，更好地近似自然语音分布，从而使效用估算能够反映匿名化语音在现实场景中的可用性和听者体验。更广泛地说，这涉及使用匿名化数据训练或微调下游模型，使其能够在原始语音和匿名化语音上都表现良好。

对于*LibriSpeech* 开发集和评估集中的每条话语，自动语音识别系统都会输出一个词序列。使用原始数据或匿名化数据估算的词错误率（WER）按照以下公式计算：

$$\text{WER} = \frac{N_{\text{sub}} + N_{\text{del}} + N_{\text{ins}}}{N_{\text{ref}}},$$

其中，N_{sub}、N_{del}和 N_{ins} 分别表示替换、删除和插入错误的数量，N_{ref} 表示参考文本中的词数。词错误率越低，语言内容的保留效果越好，因此效用也越高。

3.3.2. 非加权平均召回率（***UAR***）

情感内容的保留情况通过语音情感识别系统（记为 SER_{eval}）进行评估，该系统使用*SpeechBrain* 为*IEMOCAP* 数据库提供的配方进行训练。基于 wav2vec2 的模型针对表3所示的每个训练折分别进行训练。语音情感识别性能使用 *IEMOCAP* 开发集和评估集进行测量，采用原始或匿名化话语。我们按照以下公式计算非加权平均召回率（UAR）：

$$\text{UAR} = \frac{\sum_{i=1}^{N_{\text{class}}} R_i}{N_{\text{class}}}.$$

其中，每个类别的召回率 R_i 计算为该类别中真正例数除以该类别的样本总数，$N_{\text{class}} = 4$ 表示类别数。UAR 在五个折上取平均值。UAR 越高，情绪状态保留得越好，效用也越高。这里采用非加权平均（UAR），而非加权平均，以防高频情感类别主导该指标；鉴于*IEMOCAP*中存在类别不平衡，这一点尤为重要。

3.4. 隐私-效用权衡

与2022年版一样，本文考虑了多个评估条件（隐私类别），这些条件由一组最低目标隐私要求定义。对于每项最低目标隐私要求，满足该要求的提交系统将根据其在每项效用指标上分别产生的效用进行排名。其目标是在多个工作点衡量隐私-效用权衡，例如，当系统配置为以牺牲效用为代价提供更高隐私，反之亦然时。该评估框架更符合现实世界中的用户隐私期望，并为参赛者提供明确的优化目标，利用我们已建立的隐私与效用指标进行更全面的评估。隐私要求由 N 最低目标等错误率定义：$\{\text{EER}_1, \ldots, \text{EER}_N\}$。每个最低目标等错误率构成一个独立的评估类别。类别 i 的提交系统必须在VoicePrivacy 2024评估集上达到高于 EER_i的平均等错误率。本次挑战赛包括 $N = 4$ 个类别，分别具有

[6]https://huggingface.co/speechbrain/asr-wav2vec2-librispeech[7]
https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

最低目标等错误率为 10%、 20%、 30%和 40%。在每个隐私（等错误率）类别中，系统按照词错误率（越低越好）和非加权平均召回率（越高越好）进行排名。根据该方法得出的示例结果和系统排名如图2所示。

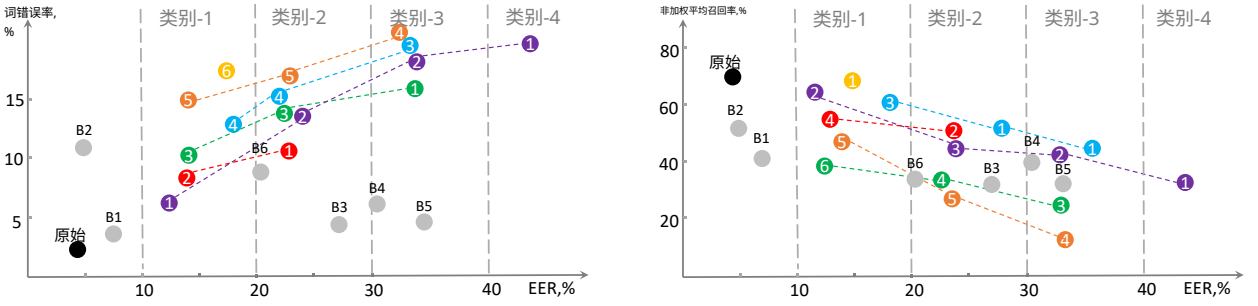


图 2： 针对 4 个最低目标等错误率，根据隐私（等错误率）和效用（词错误率和非加权平均召回率）结果得出的系统排名示例。不同颜色代表 6 个不同的团队。每个圆圈中的数字表示系统在相应类别中的排名。灰色圆圈代表基线系统，黑色圆圈代表原始（未经处理的）数据。

4. 基线语音匿名化系统

本节概述了提供给参赛者作为灵感、用于开发自身解决方案的基线语音匿名化系统。为衡量进展，我们纳入了两个在以往挑战赛届次中建立的基线系统（**B1** 和 **B2**）。我们还提供了四个新的基线系统（**B3**、**B4**、**B5**和 **B6**），这些系统在效用方面有所取舍，同时提升了匿名化性能。新的基线系统使用不同资源进行训练。下文将逐一介绍各个基线。更多详情请参见评估计划（Tomashenko 等人，2024a）以及公开提供的基线配方脚本[8]。

4.1. 使用 x 向量和神经源滤波器模型进行匿名化：B1

该**B1** 基线匿名化系统（Srivastava 等人，2020a、2022年），如图3所示，基于 VPC 2022 **B1.b** 系统，但在话语级进行匿名化。系统首先提取三种特征：用于语言内容的 256 维瓶颈（BN）特征、基频（F0）以及 512 维说话人 x 向量。BN 特征由一个在 Povey 等人，2018年）数据上训练的因子化时延神经网络（TDNN-F）模型提取，该数据包括 *LibriSpeech*（*train-clean-100*、*train-other-500*）；x 向量则来自一个在 *VoxCeleb-1*、*2*上训练的 TDNN。随后，从一个不重叠的说话人池（*LibriTTS train-other-500*）中选出的 200 个距离最远的候选 x 向量里，随机抽取 100 个并取平均，得到匿名化 x 向量。最后，使用匿名化 x 向量、BN 和 F0 特征，通过在.....上训练的神经源滤波器（NSF）模型合成匿名化语音

LibriTTS 训练集- *clean-100*采用 HiFi-GAN（Kong 等人，2020年）架构。

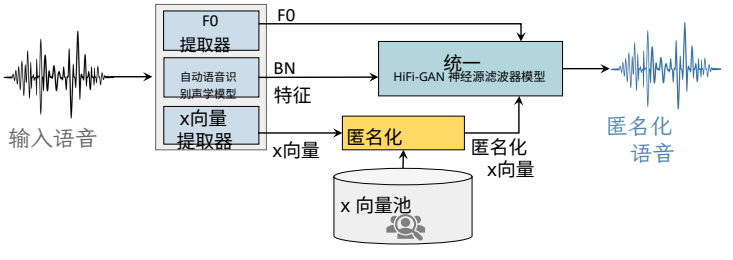


图 3： 基线匿名化系统 **B1**。

[8]2024 年 VoicePrivacy 挑战赛的匿名化基线和评估脚本:
https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2024

4.2. 使用 McAdams 系数进行匿名化：B2

该**B2** 基线匿名化系统（Patino 等人，2021年）仅使用信号处理技术，无需使用训练数据。该系统采用 McAdams 系数（McAdams，1984），对输入语音信号进行线性预测编码（LPC）分析，并移动由此得出的极点位置。逐帧进行 LPC 分析，以提取系数和残差。LPC 系数会转换为 z 平面极点位置，其中每个极点代表与某个共振峰位置对应的谱峰。McAdams 变换仅用于修改虚部非零的极点：将其相位 ϕ （0 到 π 弧度）提升至系数 α次方，从而使相位偏移 ϕ^{α}。对于每个话语， α 从 $U(0.5, 0.9)$中采样，使其围绕 $\phi = 1$ 弧度（≈20.5 kHz）收缩或扩展。共轭复极点会向相反方向移动。随后将完整的极点集合转换回 LPC 系数，并与残差结合，合成一帧匿名化语音。

4.3. 匿名 ym化：使用语音转写和生成对抗网络：B 3

基线**B3**，如图4所示，是一种语音合成系统，其条件设定旨在保留语言和一般韵律信息，同时替换说话人声音（Meyer 等人，2023a，b）。其核心是一个生成对抗网络（GAN），用于生成一个人工伪说话人嵌入（Meyer 等人，2023c）。**B3** 通过三个步骤实现语音匿名化。首先，从输入中提取说话人嵌入、语音转写、基频（F0）、能量和音素时长。说话人嵌入由一个经过调整的全局风格标记模型提取（Wang 等人，2018年），该模型使用*LibriTTS train-clean-100* 数据集训练。语音转写由一个端到端自动语音识别模型生成，该模型采用混合 CTC-Attention 架构、Branchformer（Peng 等人，2022年）编码器和 Transformer 解码器，并使用*LibriTTS train-clean-100* 和 *train-other-500* 数据集进行训练。其次，通过将原始说话人嵌入替换为使用 Wasserstein 生成对抗网络生成的人工嵌入来进行匿名化（Arjovsky 等人，2017年），并确保人工嵌入与原始嵌入之间的余弦距离

嵌入向量超过 0.3。生成对抗网络模型使用 *LibriTTS train-clean-100*、*RAVDESS* (Livingstone 和 Russo，2018年)以及*ESD* (Zhou 等人，2021年）数据集进行训练。此外，每个音素的基频和能量值都乘以从 $U(0.6, 1.4)$ 中独立均匀采样的随机值，以去除说话人特定韵律模式，同时保留整体韵律动态。每个音素的随机值均单独选取。最后，将匿名化说话人嵌入、修改后的基频和能量特征、原始音标转录以及原始音素时长输入基于 *FastSpeech2* (Ren 等人，2020年）和 HiFi-GAN（Kong 等人，2020年）的语音合成系统中，该系统由 *IMS-Toucan* 工具包（Lux 等人，2021年）实现；两者均使用 LibriTTS train-clean-100 进行训练，用于合成匿名化输出。

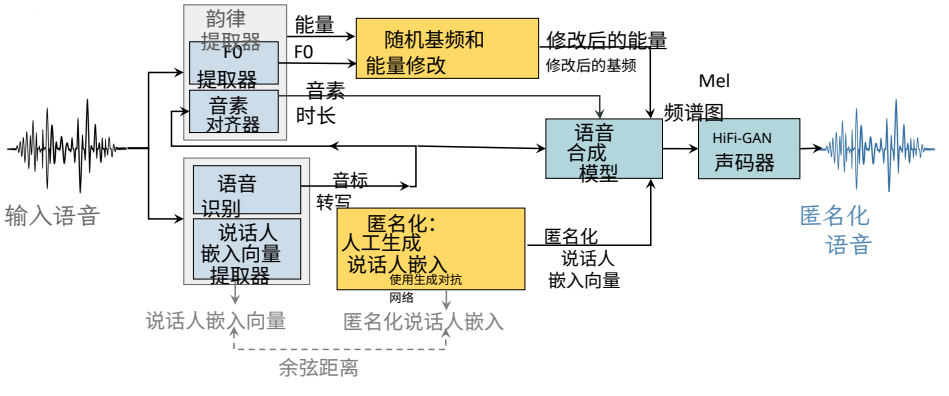


图4：基线匿名化系统**B3**。

4.4. 使用神经音频编解码器语言建模进行匿名化：B4

基线**B4** (Panariello 等人，2024a）如图5所示，其基于神经音频编解码器（NAC）语言建模技术（Borsos 等人，2023年；Wang 等人，2023年）。NAC 是一种编码器-解码器神经网络，可将输入信号编码为一系列离散的声学令牌 $\mathbf{a} \in \{1, \ldots, N_Q\}^{Q \times T_A}$，随后再解码为波形。 T_A 表示输入被分割成的时间帧数， Q 表示与每个时间帧相关联的词元数，每个词元取自 Q 个不同的词元字典。词元

是介于 1 和 N_Q 之间的整数。声学令牌被认为能够捕捉个体语音的特征。系统从一组伪说话人的语音中提取若干组声学令牌，从而获得声学提示池 $\mathbf{A}$。语义提取器用于提取离散语义令牌序列 $\mathbf{s} \in \{1, \ldots, N_S\}^{T_S}$，该序列编码所说的内容，其中 T_S 是时间帧的数量。随后将其与随机选择的声学令牌序列 $\tilde{\mathbf{a}} \in \mathbf{A}$ 连接，形成序列 $(\mathbf{s}, \tilde{\mathbf{a}})$。类似 GPT 的仅解码器 Transformer 使用该序列作为提示，生成声学令牌 $\mathbf{a}$，这些令牌反映了 $\mathbf{s}$ 中编码的语义以及 $\tilde{\mathbf{a}}$ 中编码的风格。NAC 解码器模块随后将 $\mathbf{a}$ 转换为波形，其中包含输入的语义内容，但使用的是另一位伪说话人的声音。NAC 是 EnCodec[9]（Défossez 等人，2023年），其使用 DNS 挑战赛（Dubey 等人，2022年）和 Common Voice（Ardila 等人，2020年）数据集中的语音片段进行训练，并额外使用来自 AudioSet（Gemmeke 等人，2017年）、FSD50K（Fonseca 等人，2022年）和 Jamendo（Bogdanov 等人，2019年）数据集的非语音片段。语义提取器由 HuBERT 特征提取器（Hsu 等人，2021年）和长短期记忆网络后端组成；前者使用 LibriSpeech-train-960 数据集进行训练，后者根据特征向量预测词元索引。仅解码器模型是 Bark TTS 模型公开提供的检查点。[10]

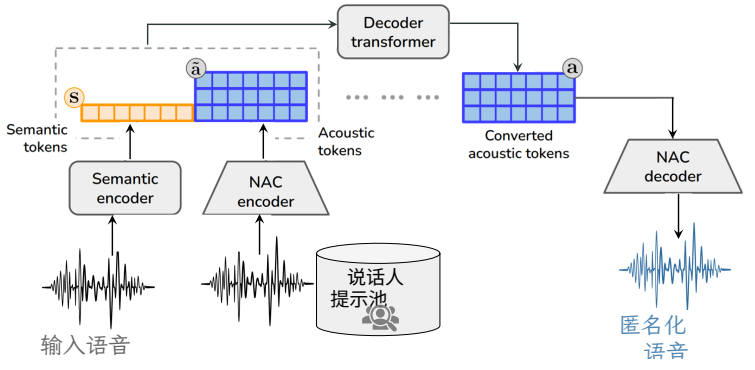


图 5： 基线匿名化系统 **B4**。

4.5. 使用带矢量量化（VQ）的 ASR-BN 进行匿名化：B5 和 B6

由**B5** 和 **B6** 基线构成的处理流程由Champion (2023年) 提出，并如图 6所示，与**B1**相似。不过，该流程通过引入向量量化（VQ）来增强语言属性与说话人属性的解耦。该流程从经过训练、用于识别左双音素的自动语音识别声学模型中提取基频（F0）和声学 VQ 瓶颈（VQ-BN）特征。随后，将 VQ-BN 特征、基频特征以及以独热向量形式表示的目标说话人表示（该向量对应训练数据集中的一名特定说话人）一同输入 HiFi-GAN 网络，以合成匿名化输出。使用两种不同的自动语音识别声学模型对应两种不同的基线： **B5** 所使用的声学模型由预训练的 wav2vec 2.0 模型和三个额外的 TDNN-F 层组成；而 **B6** 所使用的声学模型由 12 个 TDNN-F 层组成。在两种模型的最后一个 TDNN-F 层中，均在首次激活之后应用 VQ（TDNN-F 的 256 维内部瓶颈）。该过程使用有限集合中的等维向量近似连续向量。引入 VQ 可最大限度地减少瓶颈特征中编码的说话人信息。 **B5** 和 **B6** 基线采用多个预训练模型。wav2vec 2.0 模型使用 VoxPopuli[11] 数据集中的 24.1k 小时无标注多语言西日耳曼语语音进行预训练，随后使用 *LibriSpeechtrain-clean-100* 数据集进行微调。 **B6** ASR-BN 提取器以梅尔滤波器组输入特征为输入，并使用 *LibriSpeech train-other-500* 和 *train-clean-100* 数据集进行训练。最后，HiFi-GAN 模型使用 *LibriTTS-train-clean-100*进行训练，从而得到 247 个独热说话人向量。

5. 提交的匿名化系统

VPC 2024吸引了来自16个国家学术、工业和政府机构的100多名参赛者，由40个团队代表。总计，我们收到了来自13个团队的36个匿名化系统（列出

[9]https://github.com/facebookresearch/encodec[10]https://github.com/suno-ai/bark[11] https://dl.fbaipublicfiles.com/voxpopuli/models/wav2vec2_large_west_germanic_v2.pt

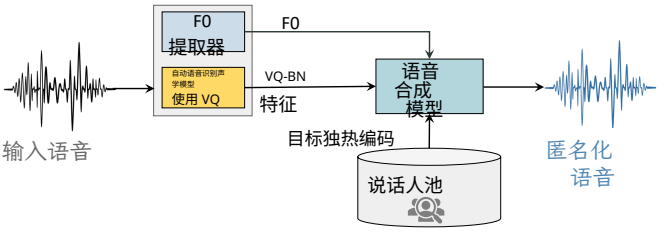


图 6： 基线匿名化系统 **B5** 和 **B6**。

表 6参赛团队。

团队名称（参考文献）	团队编号	所属机构
Anemone（Hua 等人，2024年）	T7	中国科学院声学研究所；中国科学院大学，中国— 中国科学院大学，中国
约翰斯·霍普金斯大学计算机科学与语言处理中心（Xinyuan 等人，2024年）	T8	约翰斯·霍普金斯大学，美国
LongYuan（Tan 等人，2024年）	T9	南京邮电大学；南- 京 Longyuan 信息技术有限公司，中国
NPU-NTU (Yao 等人，2024a)	T10	音频，语音 and 语言处理团队 (ASLP@NPU)；西北工业大学计算机科学学院，南洋理工大学；中国香港理工大学
NTU-NPU (Kuzmin等人，2024年)	T12	中国南洋理工大学；信息研究所通信研究所，新加坡科技研究局，新加坡；音频、语音与语言处理团队（ASLP@NPU），中国；香港理工大学，中国
ADRES（Leang等人，2024年）	T14	大学格勒诺布尔阿尔卑斯，法国国家科学研究中心。法国；研究所数字研究与创新，柬埔寨金边
南开大学人类语言技术实验室（He 等人，2024年）	T17	南开大学计算机科学学院，中国
Q（Li 等人，2024年）	T18	奇富科技；复旦大学，中国上海
DFKI_SLT（Das 等人，2024年）	T19	语音与语言技术，德国人工智能研究中心（DFKI）；柏林工业大学柏林，德国
USTC-PolyU(Gu 等人，2024年)	T25	NERC-SLIP，中国科学技术大学；香港理工大学，中国香港
V-Beam（Lee 等人，2024b，a）	T30	西江大学，首尔；梨花女子大学，首尔，韩国
KIT-ISL（Akti 等人，2024a，b）	T33	卡尔斯鲁厄理工学院（KIT），德国；卡内基梅隆大学（CMU），美国
Orange_Shiva（Le Blouch 等人，2024a，b）	T38	Orange，法国

见表 6）成功提交结果的参赛者）。每个团队提交了 1 至 6 个系统，分别针对不同的隐私类别和效用性能。参赛者开发的语音匿名化解决方案采用了多种技术方法。系统概览见表 7。不符合 VPC2024 要求的情况以红色星号（*）标示，并在正文中进一步讨论，不符合要求的原因以红色标出。

5.1. 基于神经语音转换（VC）的匿名化系统

大多数提交的匿名化系统都基于神经语音转换技术，重点在于显式属性解耦。这些系统分别提取语言内容（例如瓶颈特征，或*WavLM* 和*HuBERT*等自监督学习特征）、说话人身份（x 向量或 ECAPA 嵌入）、韵律（基频）和情感的表示。随后，在保留或修改其他属性的同时，它们使用伪说话人嵌入替换说话人特征，并通过 HiFi-GAN 等声码器合成语音信号。其变体包括 k 近邻语音转换（kNN-VC），该方法在 *WavLM* 特征空间中执行帧级最近邻替换，以及增强情感的方法。

5.1.1. k 近邻语音转换（*kNN-VC*）

两个团队，即**T8** 和**T19**，采用了 k 近邻语音转换（kNN-VC）。该方法最初由Baas 等人（2023年）提出。kNN-VC 方法不使用解耦的说话人嵌入和说话人无关表示，而是使用一个*WavLM* 模型，从输入话语中提取单一表示序列，并将其替换为从另一个说话人的话语中提取的“相似”表示；该说话人是从一个候选池中随机选出的。相似表示是原始表示的前 k 个最近表示的平均值。其假设是，相似表示编码了与输入相似的语言内容，但对应不同的说话人身份。随后，使用以所得表示为条件的波形生成器生成匿名化输出。

T8-2 基于 *WavLM* 特征的 *kNN-VC* 系统（Xinyuan 等人，2024年）：一种改进版 kNN-VC 系统，使用 *WavLM* 特征（Chen 等人，2022年）。源说话人和目标说话人的话语会被转换到*WavLM*特征空间。随后在帧级应用 kNN 回归，再使用 HiFi-GAN 声码器合成输出。T8-2 实现采用 $k = 4$ 以及使用*LibriSpeech train-clean-100*训练的声码器。为改进匿名化，系统会从*VoxCeleb*语料库中随机选取目标说话人。

T19-1 单层 *kNN-VC*（*KNNS*）（Das 等人，2024年）：基于 kNN-VC，采用单层*WavLM* 特征提取（第 6 个 Transformer 层）。源话语经过*WavLM Large* 处理，以提取帧级特征。从情感语音数据库（*ESD*）中的英语说话人里随机选取一名目标说话人。目标说话人的所有话语均由*WavLM* 处理，以构建匹配特征集。对于每个源帧，从目标说话人的特征池中找出排名前4 的最近邻，并用它们的平均值替换原始帧。预训练 HiFi-GAN 声码器会重建音频波形。

T19-2 多层 *kNN-VC*（*KNND*）（Das 等人，2024年）：通过同时使用第 6 层和第 12 层 **T19-1** 的输出进行扩展*WavLM* 。第 12 层能更好地捕捉情感线索，从而增强情感保留。由于双层特征拼接，查询和匹配集的大小翻了一倍。HiFi-GAN 声码器增加了卷积预网络模块，并针对这种双输入格式进行联合预训练。

5.1.2. 属性解耦（语言内容、说话人、基频、情感等）

T7-1,2 在模型参数层面融合语音转换说话人模型进行匿名化（Hua 等人，2024年）：通过在参数层面融合多个微调后的单目标说话人端到端语音转换模型来实现匿名化。情感通过两种不同的情感编码器进行迁移：**T7-1** 采用基于 *wav2vec 2.0* 模型的情感编码器，该模型输出的情感嵌入作为先验编码添加到声学模型中；**T7-2** 采用由多头注意力模块组成的全局风格令牌（GST）编码器（Wang 等人，2018年）。

T9 *GMM-blender* (Tan 等人，2024年)：使用 *WavLM* 模型（第 6 层输出）提取语音特征。*GMM-Blender* 方法通过对匿名化池中的特征进行插值，生成丰富多样的匿名模板。该*GMM-Blender* 首先随机选择 M 位说话人，从 *LibriSpeech* 语料库中构建动态匿名化池。对于每条待匿名化的语音话语，系统会从该池中随机选取四位说话人，每位贡献 50 个语音样本（总计 200 个样本）。*WavLM* 嵌入向量从这些样本中提取，并通过高斯混合模型（GMM）进行聚类，从而确定 2000 个聚类中心。在每个聚类中心，选取距离最近的四个特征嵌入，并使用随机权重将其混合，生成复合匿名化嵌入向量模板。为在匿名化过程中保留情感特征，系统还采用了基于 wav2vec2 的情感特征提取器（*wav2vec2-large-robust-12-ft-emotion-msp-dim* (Wagner 等人，2023年)）。从源语音中提取的情感特征与匿名化语音特征相融合。这种融合表示

随后提供给经过修改的 HiFi-GAN 声码器以合成波形。训练目标将标准的 HiFi-GAN对抗损失与情感相似性项相结合，约束合成语音保持原始语音的情感特征。

***T12-5 B5*-基础的采用均值回归基频平滑的匿名化 (Kuzmin等人，2024年)**：该系统通过应用“均值回归”机制修改 **B5** 基线中的基频（F0），即将每一帧的 F0 估计值与相邻帧的移动平均值进行插值。这会降低音高变化，并平滑音高轮廓，从而有助于实现匿名化。

***T12-6 B5*-基础的采用均值回归基频平滑和加性高斯噪声的匿名化 (Kuzmin等人，2024年)**：与 **T12-5**类似，该系统还会在推断过程中，以受控功率将加性白高斯噪声（AWGN）显式应用于经过均值回归处理的 F0。这通过增加说话人识别的难度进一步增强隐私，同时保持合理的语音效用。

***T14-1*、*2* 带韵律参数的*VQ*变分自编码器（*VAE*）（Leang等人，2024年）**：使用定制的卷积模块从输入语句中提取内容特征，再通过VQ量化这些特征，以将残留说话人信息与内容信息解耦。使用YAAPT估计输入语句的对数归一化F0曲线（Kasi和Zahorian，2002），然后将其输入双向门控循环单元（GRU）模型（Cho等人，2014）。声音身份通过ECAPA-TDNN嵌入向量编码，并使用广为人知的基于池的匿名化函数进行匿名化（类似于**B1**）。使用HiFi-GAN合成输出。**T14-1**使用从输入中提取的标准归一化F0和能量韵律特征。**T14-2** 在合成过程中引入F0随机缩放，进一步模糊说话人身份。缩放使用的随机因子范围为 [0.8, 1.2]。其余架构与**T14-1**保持一致。

***T18-1*、*2* 基于 *FreeVC* 的情感增强匿名化 (Li 等人，2024年)**：这两个系统均基于 FreeVC（Li 等人，2023年）语音转换框架，该框架采用经生成对抗网络训练增强的条件变分自编码器（CVAE），无需文本标注即可实现单样本语音转换。这两个系统都集成了情感特征提取，以情感信息为条件对解码器进行调控，旨在匿名化过程中保留情感内容。**T18-1** 采用在*MSP-Podcast* 数据集上微调的 wav2vec 2.0 large-robust模型作为情感编码器。以情感信息为条件后，该系统的工作方式与*FreeVC*相同。对于**T18-2**，使用 *openSMILE* (Eyben 等人，2010)提取的手工特征经长短期记忆网络处理后，生成情感嵌入。

***T19-3* 基于矢量量化互信息的语音转换（*VMC*）(Das 等人，2024年)**：该系统基于矢量量化和基于互信息的语音转换（VQMIVC）方法（Wang 等人，2021年）。内容提取器采用矢量量化（VQ），并在量化内容特征上使用基于对比预测编码的训练损失（Wang 等人，2021年）。训练准则旨在最小化内容、说话人和基频表示之间的互信息。内容提取器和说话人提取器均为卷积深度神经网络，二者从头开始联合训练。基频提取器基于*WORLD*，而波形生成器则采用独立训练的并行 WaveGAN。

T25-1,2* 和***T25-p1,p2* 通过内容与非内容解耦及 *GST* 建模实现情感感知匿名化 (Gu 等人，2024年)**：两个系统采用了相同的通用框架，包括：(1) 一个基于基线系统**B5** 改编的内容编码器，使用通过 wav2vec 2.0 提取的 ASR-BN 特征，并结合向量量化（VQ），生成离散内容表示；(2) 一个利用 GST 框架对说话人身份、韵律和情感等非内容属性进行建模的非内容学习器，该学习器通过参考编码器和注意力模块实现；(3) 一个 HiFi-GAN 声码器，根据内容与非内容嵌入的组合合成波形；(4) 一个情感识别模块，用于预测输入语音的情感，并从语音池中选择情感匹配的参考话语。匿名化流程通过对原始语音的内容和情感进行编码，选择情感相同的参考话语以提取伪说话人嵌入，再结合原始内容与伪非内容信息合成匿名化语音。两个提交的系统主要区别在于非内容学习器所用情感标注语料库的组成：在**T25-1**中，*LibriTTS train-clean-100* 数据集补充了来自*ESD*的英语语音；在**T25-2**中，仅使用了*ESD* 部分。测试时使用在*IEMOCAP*上微调的语音情感识别（SER）来选择情感匹配的话语。然而，该团队在评估后的分析表明，违规并不严重。具体而言，[illegible]

采用另一种语音情感识别模型：多模态多任务学习模型进行评估。这是一种用于语音情感识别的多模态多任务学习方法（Ghosh 等人，2023年）。该多模态多任务学习模型在*MSP-Podcast* 语料库上预训练，并在*ESD* 数据集上微调。相应的系统**T25-p1** 和**T25-p2**以多模态多任务学习模型取代*IEMOCAP* 语音情感识别，符合挑战赛规则。**T25-p1** （类似于**T25-1**）使用*ESD* 作为参考池，而**T25-p2** 则同时使用*ESD*和*LibriTTS train-clean-100*。 **T25-p1** 和 **T25-p2** 的结果表明，多模态多任务学习模型提升了非加权平均召回率，且未显著降低其他指标（见表9）。

T33-1、***T33-2*** 采用 *HuBERT* 和自动语音识别模型的基于 *FreeVC* 的语音转换系统（Akti 等人，2024a，b）：一种基于*FreeVC* （Li 等人，2023年）的零样本语音转换系统：从基于 Transformer 的自动语音识别系统的最后一层提取内容特征，并与目标说话人嵌入向量和基频估计值结合，然后生成波形。目标说话人嵌入向量从 VCTK 训练集中随机选取。**T33-1** 与**T33-2** 类似，但使用预训练的*HuBERTBase* 模型，而非基于 Transformer 的自动语音识别系统。使用 HuBERT 第 9 个 Transformer 块的输出作为内容特征。

T38-1、***T38-2***、***T38-3***、***T38-4***、***T38-5*** 基于 *DISSC* 的系统 (Le Blouch 等人，2024a，b）：这些系统都基于离散单元说话风格转换（DISSC）（Maimon 和 Adi，2023a）这一语音转换方法，该方法使用解耦的自监督学习表征。**T38-1** 是原始的 DISSC 系统。通过 k 均值量化、用于编码语言内容的 HuBERT 离散表征被输入预测器，以估计每个离散单元的时长和基频。HiFi-GAN 使用内容特征和从 VCTK 训练集中随机选取的目标说话人嵌入向量来合成输出。**T38-2** 使用从输入波形中提取的时长和基频。为扩展同一系统并保留表现力，**T38-3** 使用通过 *MSP-Podcast* 训练集微调的 HiFi-GAN。**T38-4** 对基频估计值添加随机扰动。**T38-5**是 **T38-1**的变体，使用 *MSP-Podcast* 数据微调时长和基频预测模块。最后，**T38-6** 在 **T38-5** 的基础上，将说话人嵌入向量与使用预训练自监督学习模型从源波形中提取的情感嵌入向量拼接起来。

5.2. 基于神经音频编解码器的匿名化系统

基于神经编解码器的语音匿名化系统代表了一种利用神经音频编解码器（NAC）强大表示能力的现代方法。这些编解码器将语音转换为量化编码，在捕捉语言内容和韵律特征的同时，有效地对说话人相关信息形成瓶颈。通过将语音编码为离散的瓶颈表示，这些系统能够天然地解耦说话人身份、语言内容和韵律。匿名化通过在瓶颈层面操纵或替换说话人身份成分来实现，同时保留内容和韵律线索，以维持语音自然度和可懂度。与大多数依赖显式特征提取和修改的传统语音转换方法相比，神经编解码器系统在统一的编解码器框架内运行，能够自然地分离并重新组合语音成分。

T10-1,2* 采用串行解耦和多级蒸馏的解耦神经编解码器匿名化（Yao 等，2024a）：这是一对神经编解码器系统，使用说话人身份、语言内容和基频的解耦表示。系统通过帧级情感蒸馏保留情感线索，同时通过语义教师模型和自监督说话人蒸馏，将语言内容与说话人信息解耦。匿名化通过加权求和实现：将从说话人池中得出的平均声音身份与从高斯分布中随机选取的身份相结合。匿名化系统中的参数 α 是一个权重，用于控制说话人身份匿名化过程中平均说话人身份向量与随机采样的说话人身份向量之间的混合比例：$\mathbf{s}_{\text{anon}} = \alpha \cdot \bar{\mathbf{s}} + (1-\alpha) \cdot \hat{\mathbf{s}}$，其中：$\mathbf{s}_{\text{anon}}$ 是匿名化说话人身份向量，$\bar{\mathbf{s}}$ 是根据选定的一组候选说话人身份计算得出的平均说话人身份向量，$\hat{\mathbf{s}}$ 是随机生成的说话人身份向量，$\alpha \in [0,1]$ 是控制平均向量和随机向量影响程度的混合权重。较高的 α 会赋予平均说话人身份更大的权重，通常会降低匿名性，但更有利于效用保留；较低的 α 则会增加随机性，增强匿名性，但可能降低效用。参数 α 可用于调整隐私类别。两个系统的权重不同，因而属于不同的隐私类别：T10-1 的 $\alpha = 0.9$ 对应第三类隐私类别（等错误率为 $\geq 30\%$），T10-2 的 $\alpha = 0.8$ 则属于最后一类（等错误率为 $\geq 40\%$）。这些系统未能完全满足挑战赛对自动说话人验证模型训练所用训练数据进行匿名化的要求。具体而言，训练数据应在话语级进行匿名化，并采用与测试

T12-1 基于 *FACodec* 的匿名化：采用池化平均、跨性别转换和高斯噪声 （Kuzmin 等人，2024年）：使用因子化神经语音编解码器（FACodec）（Ju 等人，2024年）对韵律、内容和其他声学细节进行解耦。该编解码器以音色提取器生成的说话人嵌入作为条件。匿名化通过将编解码器提取的原始说话人嵌入替换为匿名化的说话人嵌入来实现。所采用的匿名化策略包括：基于池化的平均、跨性别转换，以及向说话人嵌入添加具有不同噪声功率水平的白高斯噪声（AWGN）。

T17* 基于 *FACodec* 的匿名化：增强的韵律转换 （He 等人，2024年）：在 FACodec 中加入额外的韵律匿名化模块，该模块以梅尔频谱图和目标说话人嵌入为输入。利用与梯度反转层配对的说话人嵌入提取器，抑制使用韵律解码器生成的潜在韵律特征中包含的说话人特有特征。韵律解码器生成包含目标说话人韵律信息的新梅尔频谱图。还会将从原始话语中提取的音色嵌入替换为目标说话人的音色嵌入。在匿名化之前，基于整个数据集计算说话人嵌入的中心，而这些嵌入由 FACodec 音

5.3. 级联自动语音识别+文本转语音匿名化系统

本节介绍采用自动语音识别后接文本转语音合成这一串行流程的匿名化技术。这些级联系统会将语音转写为词语或音素序列，作为中间表示，再据此重新合成，生成匿名化语音输出。

5.3.1. 传统级联自动语音识别-文本转语音

这些系统采用传统方法，将自动语音识别生成的文本转写结果直接传递给文本转语音系统，而不显式保留韵律或副语言属性。

T8-1 级联自动语音识别（*Whisper*）与文本转语音（*VITS*）（Xinyuan 等人，2024年）：使用中型英语*Whisper* 模型从输入中生成文本转写结果。随后使用基于Kim 等人，2021年）训练的 VITS 文本转语音模型，*LibriTTS* 数据集生成匿名化输出。

5.3.2. 保留副语言属性的级联自动语音识别-文本转语音系统

这些先进的级联系统整合了保留情感、语调和说话风格等副语言特征以及语音内容的机制。

T12-2*、*T12-3 ***B3***基础系统，集成了情感嵌入，并采用跨性别平均和选择进行匿名化（Kuzmin等人，2024年）：两种系统均基于**B3** 基线系统，并采用*FastSpeech2* 模型和 HiFi-GAN 声码器合成语音。它们使用经过微调的*wav2vec2.0*模型提取情感嵌入，以增强情感保留。两种系统的说话人匿名化均采用跨性别相关策略。它们的差异在于说话人匿名化机制、嵌入向量的使用方式以及韵律修改模块。系统**T12-2** 通过从源说话人的 200 个距离最远的嵌入向量中随机选取 100 个并求平均，再结合跨性别选择来实现说话人匿名化。系统**T12-3** 则按话语随机从跨性别集合中选择说话人，而非对嵌入向量求平均。此外，**T12-3** 通过对基频和能量特征进行归一化，并使用范围为 [0.7, 1.3]。

T12-4 ***B3***-基于跨性别选择式匿名化 (Kuzmin 等人，2024年）：该系统基于**B3** 基线，但不会修改韵律。该系统不使用生成对抗网络生成的伪说话人嵌入，而是采用跨性别选择式匿名化来实现说话人嵌入匿名化。

T30-1 ***B3***-基于嵌入池扰动和 k 均值聚类的匿名化 (Lee 等人，2024b，a)：B3 的生成对抗网络说话人嵌入生成过程被一种基于池的匿名化算法取代。选取池中与输入说话人嵌入余弦距离最大的说话人嵌入。随后，随机选取一些嵌入维度，并用从 500 个池聚类质心中随机选取的说话人嵌入所对应的维度替换。

T30-2 级联自动语音识别+融合情感增强特征的文本转语音匿名化 （Lee 等人，2024b，a）：将 Whisper 自动语音识别与结合对抗学习的变分推断用于端到端文本转语音（VITS）（Kim 等人，2021年）文本转语音系统。输入波形中提取的基频估计值被转换为潜在特征，随后通过交叉注意力与一组情感原型向量相结合。组合后的特征被输入逐特征线性调制（FiLM）层，其输出与 Whisper 自动语音识别文本嵌入相乘。所得特征与 VITS 一起用于波形生成。

5.4. 混合匿名化系统

混合系统结合多种匿名化技术，以平衡隐私与效用。不同方法呈现出不同的权衡：级联自动语音识别+文本转语音可实现较强的匿名化，但会降低情感保留效果；而语音转换能够保留副语言内容，却会牺牲部分隐私。混合系统通过支持动态方法选择来解决这一问题，例如挑战赛提交系统中提出的随机混合方法，从而实现对隐私—效用权衡的细粒度控制。

T8-3、***T8-4***、***T8-5*** 混合 (Xinyuan 等人，2024年)：随机选择以下两种方法之一进行匿名化。第一种是级联 ASR-TTS 系统 **T8-1** ，第二种是 kNN-VC 系统 **T8-2**。这种混合方法在匿名化强度与效用保留之间取得平衡： **T8-1**系统能更有效地抑制语音信息，但效用有所降低；而 **T8-2** 系统能更好地保留语言和情感内容，但匿名化程度有所降低。

表 7：提交的系统概览。第一列中的颜色标示相应系统所属的隐私类别：等错误率≥ 10%，等错误率≥ 20%，等错误率≥ 30%，等错误率≥ 40%。不合 VPC 2024 要求的项目以 示。

ID	编号（论文）	训练资源	描述
T7-1	ppg-w2vF0-fusion	LibriTTS train-clean-360, LibriTTS train-other-500, LJSpeech、ESD, 在 MSP- 上进行 wav2vec 2.0 微调播客	匿名化：在参数层面对伪说话人（多个预训练的端到端使用单说话人数据微调的语音转换模型）；情感编码器：基于 wav2vec 2.0；基频：YAAPT；语言内容：来自混-合式 Transformer-CTC 自动语音识别的音素后验概率特征
T7-2	ppg-gstF0 融合	LibriTTS train-clean-360 ，LibriTTS train-other-500, LJSpeech、ESD	与 **T7-1** 相似，但情感编码器除外：GST
T8-1	Whisper-VITS语音合成	LibriTTS ，中型英语Whisper	级联自动语音识别（Whisper）和文本转语音（VITS）
T8-2	kNN-VC	LibriSpeech train-clean-100，VoxCeleb；WavLM	基于 WavLM 特征的 kNN-VC 系统
T8-3	混合（$p = 0.2$）	如 **{T8-1 ∪T8-2}**	随机选择以下两种方法之一，用于每个话语（第二种方法的概率为 p 方法）：**T8-1** （级联自动语音识别与语音合成）和 **T8-2**（kNN-VC）
T8-4	混合（$p = 0.325$），如 **{T8-1 ∪T8-2}**		
T8-5	混合（$p = 0.4$）		
T9	GMM-Blender	ESD, LibriSpeech train-clean-360 ， train-other-500 ，WavLM, wav2vec2-large-robust-12-ft-emotion-msp-dim	匿名化：基于*GMM-Blender* 的 GMM 说话人聚类，用于生成匿名模板；情感编码器：wav2vec2-large-robust-12-ft-情感-msp-dim
T10-1*	C3	LibriSpeech, LibriTTS; WavLM	神经音频编解码器，具有特定的解耦-策略；匿名化：.....的加权和平均说话人身份（来自说话人池中的说话人）与随机生成的.....）

表 7：提交的系统概览。第一列中的颜色标示相应系统所属的隐私类别：等错误率≥ 10%，等错误率≥ 20%，等错误率≥ 30%，等错误率≥ 40%。不符合 VPC 2024 要求的项目以斜体标示。

ID	编号（论文）	训练资源	描述
T10-2*	C4		与**T10-1**类似，但对以下内容赋予较小权重：平均说话人身份
T12-1	1a	LibriTTS train-clean-100，LibriLight 训练集，NaturalSpeech3 FACodec	依赖 FACodec，匿名化：基于池化、跨性别选择-tion，添加白高斯噪声。
T12-2	1b	LibriTTS train-clean-100，train-other-500，MSP-Podcast；wav2vec 2.0	修改 **B3**；情感编码器：基于 wav2vec；匿名化：基于池化、选择+平均，余弦距离、跨性别）
T12-3	2a	LibriTTS train-clean-100，train-other-500，MSP-Podcast；wav2vec 2.0	修改 **B3**；情感编码器：微调 wav2vec 2.0 韵律：基频和能量按数值进行乘法运算通过随机值；匿名化：随机选择，跨性别
T12-4	2b	LibriTTS train-clean-100，train-other-500	修改 **B3**；匿名化：随机选择，跨性别
T12-5	3	VoxPopuli，LibriSpeech train-clean-100，LibriTTS train-clean-100	修改 **B5**；韵律：将基频曲线替换为曲线本身与其移动平均值之间的加权平均值及其移动平均。
T12-6	4		类似于**T12-5**，向……添加白高斯噪声生成的基频曲线添加白高斯噪声。
T14-1	v1	LibriSpeech，CREMA-D	矢量量化变分自编码器；韵律：归一化的基频和能量来自输入；匿名化：选择+平均 与 **B1** 相同
T14-2	v2		类似于 **T14-1**，使用随机因子进行基频缩放在范围 [0.8, 1.2] 内。
T17*	FACodec+PANO	LibriLight，LibriSpeech train-other-500；VITS（用于韵律编码器）；NaturalSpeech3编解码器	为 FA- 添加韵律匿名化模块编解码器。
T18-1	EESA(Wv)	MSP-Podcast，RAVDESS，VCTK；FreeVC、WavLM，wav2vec 2.0，wav2vec2-large-robust12-ft-emotion-msp-dim	FreeVC 与情感嵌入集成生成-erated by a fine-tuned version of wav2vec 2.0。
T18-2	EESA(OS)	MSP-Podcast，RAVDESS；VCTK、FreeVC、WavLM、wav2vec 2.0	采用情感嵌入增强的 FreeVC，通过在 openSMILE MILE 特征
T19-1	KNNS	LibriSpeech train-clean-360，ESD，RAVDESS，CREMA-D	采用 WavLM 特征的 kNN-VC 系统（取自第 6 个 Transformer 块）
T19-2	KNND		采用 WavLM 特征的 kNN-VC 系统（拼接自第 6 个和第 12 个 Transformer 块）

表 7：提交的系统摘要。第一列中的颜色标示相应系统的隐私类别：等错误率≥ 10%，等错误率≥ 20%，等错误率≥ 30%，等错误率≥ 40%。不符合 VPC 2024 要求的项目以*标示。

ID	ID（论文）	训练资源	描述
T19-3	VMC		匿名化：随机选择的说话人向量 来自 ESD 说话人；基频：PyWORLD； 语言内容：卷积神经网络 +向量量化。
T25-1*	大型： ESD+LibriTTS	ESD, LibriTTS train-clean-100, IEMO ）； wav2vec 2.0 来自**B5**的内容编码器	内容解耦（与**B5**中相同的 VQ-BN） 以及风格特征；情感迁移来自目标 说话人语音；非内容学习器：训练于 补充了 LibriTTS train-clean-100 数据集 以及来自 ESD 的英语语音；
T25-2*	小型：仅使用 ESD		与 **T25-1** 类似 ，但非内容学习器是 仅在 ESD 子集上训练
T30-1	V-Beam_sttts-tj- 方法3	LibriTTS ，RAVDESS ，ESD	修改**B3**；匿名化：从 通过 k 均值聚类创建的池中进行选择
T30-2	V-Beam方法2_	LJSpeech；Whisper、VITS	级联自动语音识别（Whisper） + 文本转语音（使用 VITS LJspeech）并整合情感类别原型 特征与基频
T33-1	freevc-hubert-base- ss	VCTK; HuBERT Base L （km500量化器）	匿名化：从 VCTK 中随机选取； 基频：未进行显式提取或转换； 语言内容：从第 9 个 HuBERT 块中提取的特征
T33-2	freevc-asr- bottleneck-f0-ss	VCTK, LibriSpeech	类似于 **T33-1** ，但内容特征是从 基于 CTC-Transformer 的最终层提取的 ASR
T38-1	M0 - 原始 DISSC	DISSC 基于.....训练的模型 VCTK	DISSC 使用说话人嵌入随机选取 se- 从 VCTK 说话人中选取的
T38-2	M1 - 韵律保留 存	与 **T38-1** 相同	**T38-1** 但使用输入波形的时长和基频—— form
T38-3	M2 - 用于情感表现力的数据集 表现力	如 **T38-1** + MSP-Podcast	**T38-1** +在 MSP-Podcast 上训练的 HiFi-GAN
T38-4	M3 - 添加随机 对韵律的敏感性		**T38-3** +基频中的随机噪声
T38-5	M4 - 韵律 预测 with MSP-Podcast		**T38-1** +在 MSP- 上训练的韵律预测器 Podcast
T38-6	M5 - 情感嵌入 向量	MSP-Podcast；DISSC模型 在 VCTK 上训练， wav2vec2-large-robust-12- ftemotion-msp-dim	**T38-5** +情感嵌入

5.5. 趋势

5.5.1. 匿名化

级联方法： 像**T8-1** 这样采用自动语音识别+文本转语音级联的系统，可实现近乎完美的匿名化，等错误率约为 48%。然而，虽然语言内容大体得以保留，但仍存在明显的折衷

在情感保留方面。为实现更好的平衡，采用了混合 策略，随机选择应用于每个话语的具体匿名化方法，从而灵活权衡隐私与效用。这种方法的优势体现在混合系统**T8-3**、**T8-4**和**T8-5** 的结果中：这些系统会随机对每个话语应用级联自动语音识别-文本转语音和 kNN-VC 方法。

说话人嵌入匿名化方法：

- 加权平均与随机混合： **T10-1、2** 系统将从说话人池中得出的平均说话人身份，与从高斯分布中采样得到的随机说话人身份相结合。混合权重参数用于控制平均成分与随机成分之间的平衡。较高的权重更倾向于采用平均身份，因此能更好地保留效用；较低的权重则会增加随机性，从而增强隐私保护。

- 基于池的选择（可选平均与跨性别转换）： 多个系统采用基于池的说话人嵌入选择策略，并在平均程度和跨性别约束方面各有不同：

  –基于平均的策略：**T12-1** 采用基于 FACodec 的系统，通过基于池的说话人嵌入平均，结合跨性别说话人选择，并向说话人嵌入注入不同噪声功率水平的加性白高斯噪声。**T14-1、2** 采用与基线**B1**相同的选择与平均方法。与**B1**类似，**T12-2** 也从一个包含 200 个嵌入向量的池中，依据余弦距离选出距离最远的 100 个嵌入向量并求平均，同时进行跨性别选择。

  – 单说话人选择策略： **T12-3** 按话语随机从跨性别说话人池中选择说话人。**T30-1** 采用基于选择的匿名化，从通过对说话人嵌入进行 k 均值聚类所创建的池中进行选择。**T33-1、T33-2** 从 VCTK 说话人语料库中随机选择单个伪说话人嵌入。**T38-1、T38-2、T38-3、T38-4、T38-5、6** 采用从 VCTK 说话人中随机选取的说话人嵌入来实现 DISSC。

- 情感匹配的伪说话人选择： **T25-1,2** 系统采用情感感知匿名化策略，从带情感标签的池中选择伪说话人。情感识别模块用于估计输入语音的情感，以便从池中选取情感匹配的参考话语。该策略有助于在替换声音身份的同时更好地保持情感一致性。

用于韵律改变的基频操控： **T12-3** 通过按值对基频和能量应用随机乘数进行缩放来改变韵律。**T12-5,6** 通过均值回归式音高平滑实现韵律匿名化，即将基频估计值与相邻帧的移动平均值进行插值。**T12-6** 还会对平滑后的基频估计值添加加性白高斯噪声。**T14-2** 在合成过程中引入 0.8–1.2 范围内的随机基频缩放，以掩盖与音高相关的说话人线索。**T17**集成了一个额外的韵律匿名化模块。

5.5.2. 情感保留

VPC 2024年的提交方案探索了多种策略，以改进语音匿名化过程中情感线索的保留。

- 预训练的基于自监督学习的情感编码器： **T7-1** 和 **T18-1** 采用在情感语料库（MSP-Podcast、ESD）上微调的 *wav2vec 2.0* 模型，提取情感嵌入，并在合成过程中以此对解码器进行条件控制。**T19-2**将单层 *WavLM* 特征（第 6 个 Transformer 块）扩展为第 6 个和第 12 个块的双层拼接，其中第 12 层能更有效地捕捉情感线索。

- 基于手工特征的情感编码：**T18-2** 使用 *openSMILE* 提取声学特征，再通过长短期记忆网络处理这些特征，以生成情感嵌入向量。

- 全局风格令牌（*GST*）方法：**T7-2** 和 **T25-1,2** 利用由多头注意力模块组成的 GST 框架，将情感和风格特征建模为可学习的风格嵌入，并通过参考编码器提取这些嵌入。

- 情感蒸馏：**T10** 的神经编解码器系统将帧级情感蒸馏作为多阶段解耦策略的一部分，明确保留情感信息，同时保留语言内容和说话人身份。

- 情感识别引导的参考语句选择：**T25-1,2** 使用在 *IEMOCAP* 上训练的语音情感识别模型来估计输入语音中的情感。随后，从情感标注语料池（*ESD*或 *ESD+LibriTTS*）中选择情感标签相匹配的参考话语，以更好地保留情感表现力。

5.5.3. 语言内容保留

内容保留策略因匿名化方法而异。

- 矢量量化（*VQ*）瓶颈特征：多个系统（**T14-1,2**、**T19-3**、**T12-5,6**和 **T25-1,2**）采用基于矢量量化的解耦，将说话人身份与语言内容分离。**T14-1,2** 使用卷积模块提取内容特征，随后通过 VQ 进行量化，以去除残留说话人信息，同时保留内容表示。与基线 **B5** 和 **B6**类似，**T12-5,6** 和 **T25-1,2** 将 VQ 应用于 ASR-BN 特征。连续瓶颈表示被离散化为词元，以减少说话人信息泄漏。这些方法的基本原理是，将内容量化为离散词元可形成瓶颈，在保留语音识别与合成所需的离散语言内容的同时，削弱说话人特异性变化。

- 预训练自监督学习表示：多个系统利用预训练自监督语音模型作为内容编码器，借助其在多种声学条件下稳健捕捉语言信息的能力。**T33-1**从*HuBERT Base* 模型中提取内容特征；**T8-2** 和**T19-1,2** 采用 *WavLM* 特征。**T8-2** 和**T19-1,2** 使用*WavLM* 特征。

- 语音转写：**T12-2,3,4**、**T30-2** 采用级联自动语音识别系统（*Whisper*、混合 CTC-注意力自动语音识别）获取语音转写或中间表示，再将其输入文本转语音系统，以鼓励在转写层面保留语言信息。

6. 结果

图7 展示了提交系统（橙色星形标记，见第5 节）和基线系统（蓝色三角形标记，见第3.4节）在效用（词错误率百分比或非加权平均召回率百分比）与隐私（等错误率百分比）方面的结果汇总。所有情况下，等错误率结果均基于组织方使用参赛者提供的匿名化数据重新训练的自动说话人验证模型得出。原始未保护数据的结果以黑色圆圈表示。每项等错误率估计值均为分别基于男性和女性说话人子集计算出的等错误率的平均值，所用数据来自LibriSpeech 测试集（干净语音）。开发集和测试数据集的详细结果见表9。

6.1. 各类别的隐私达成情况

图 8 展示了四种不同隐私阈值（等错误率的 10%、20%、30% 和 40%）下，效用指标（词错误率，% 和非加权平均召回率，%）相对于等错误率的结果与排名。词错误率数值附有 95% 置信区间。原始未保护测试数据的等错误率为 4.6%。第一类（等错误率 $\geq$ 10%）包含 28 个提交的系统和 4 个基线。其中，21 个提交的系统和全部 4 个基线满足第二类（等错误率 $\geq$ 20%）的要求。第三类（等错误率 $\geq$ 30%）包含 15 个提交的系统和 2 个基线，而只有 6 个提交的系统（**T10-2**、**T8-5**、**T25-1**、**T12-5**、**T30-2**和 **T8-1**）符合最强的第四类（等错误率 $>$ 40%）的标准。然而，其中两个（**T10-2** 和 **T12-5**）不符合挑战赛规则。级联自动语音识别与语音合成系统 **T8-1**实现了最高的等错误率，接近 50%。

6.2. 按系统类型比较性能*6.2.1.* 基于神经编解码器的系统（*T10*、*T12-1*、*T17*）

图8 显示，在最高隐私类别中，一些神经编解码器系统的表现很强。**T10-2** (等错误率=40.4%），该系统将编解码器与串行解耦和帧级情感蒸馏相结合，在隐私保护（等错误率 $\geq$ 40%）以及语言内容（词错误率=3.2%）和情感表现力（非加权平均召回率=60.8%）保留方面，实现了最佳的整体平衡。 **T17**是一个基于 FACodec 的系统，属于第二类（等错误率=27%），并且在保持效用方面也表现出色：词错误率=3%，非加权平均召回率=51.4%）。 **T12-1**采用基于池化的平均、跨性别转换和加性白高斯噪声与 FACodec 相结合，隐私保护效果较弱（等错误率=9.7%），效用表现中等。

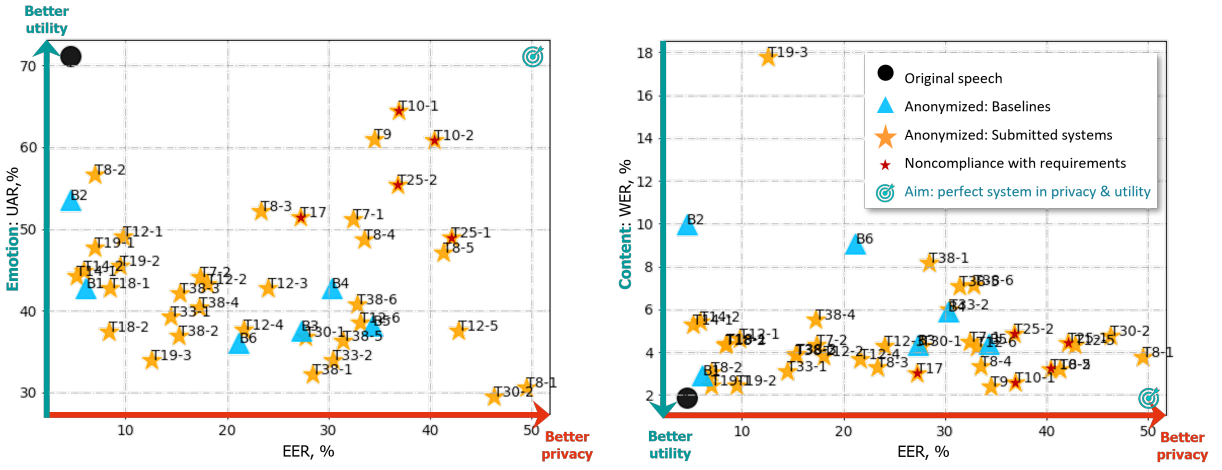


图 7： 测试数据上的挑战赛结果

6.2.2. 级联自动语音识别+文本转语音系统（***T8-1***、***T12-2***、***3***、***4***、***T30-2***）

这些系统实现了近乎完美的说话人匿名化，其中一个系统（**T8-1**）的等错误率接近 50%。然而，如图7所示，与神经编解码器和基于语音转换的替代方案相比，级联系统的非加权平均召回率始终显著偏低。 **T8-1**是一个使用 *Whisper-VITS* 的级联自动语音识别 - 文本转语音系统，隐私保护效果最强（等错误率 =49.5% ），但情感保留效果却最差的系统之一（非加权平均召回率=30.6%）。不过，其语言内容保留效果依然出色（词错误率=3.75%，在最高隐私类别中排名第三）。这种通过转录瓶颈保留内容、却导致副语言属性退化的不对称现象，凸显了基于文本的匿名化的主要局限。

在具有副语言信息保留能力的级联 ASR-TTS 系统中，处于同一最高类别的**T30-2** （等错误率为=46.3%）在两项效用指标（词错误率为=4.78%、非加权平均召回率为=29.4%）上的结果均劣于**T8-1** ，这表明，加入基于原始基频和情感原型向量的潜在特征，可能会导致隐私保护能力略有下降，同时并不能改善情感保留效果。**T12-2、3、4** ——基于**B3**的匿名化系统——分别属于第一类（**T12-2**）和第二类（**T12-3、4**）隐私类别。**T12-2** 和**T12-3** 使用情感编码器，可将非加权平均召回率提高至最高 43%，这是此类系统的最佳结果。相比之下，不使用情感编码器的类似系统（即**T12-4、B3**）的非加权平均召回率均不超过 38%。

6.2.3. 混合掺合系统（***T8-3***、***T8-4***、***T8-5***）

T8的混合方法会逐条话语随机选择级联自动语音识别+文本转语音（**T8-1**）或 kNN-VC（**T8-2**），在各项隐私-效用目标上的表现均高于平均水平。混合概率可灵活调整各隐私类别中的隐私-效用权衡：非加权平均召回率随等错误率成比例变化，而词错误率在所有系统中相近（约为 3.2–3.3%）。**T8-5**名列顶尖系统之列，其等错误率为=41.3%，同时非加权平均召回率更接近纯语音转换系统的水平（非加权平均召回率=47 %），表明混合方法能够有效实现依赖上下文的隐私效用优化。

6.2.4. 基于语音转换的系统（***T7***、***T8-2***、***T9***、***T12-5***、***6***、***T14***、***T18***、***T19-1***、***2***、***3***、***T25***、***T33***、***T38***）

纯语音转换方法通常难以达到最高隐私类别，但在某些情况下能实现良好的效用保留。kNN-VC 变体（**T8-2、T19-1,2**）的词错误率和非加权平均召回率相对较高，但其等错误率通常不超过 10%。与大多数基线语音转换系统相比，情感感知语音转换系统（如**T25-1,2** 和 **T9** ）展现出更好的情感保留能力，但在最高隐私类别中仍逊于神经编解码器性能（例如**T10-1,2**）。**T9** 在前三个隐私类别中名列前茅（等错误率=34.5 %，词错误率=2.4%，非加权平均召回率=61%）。

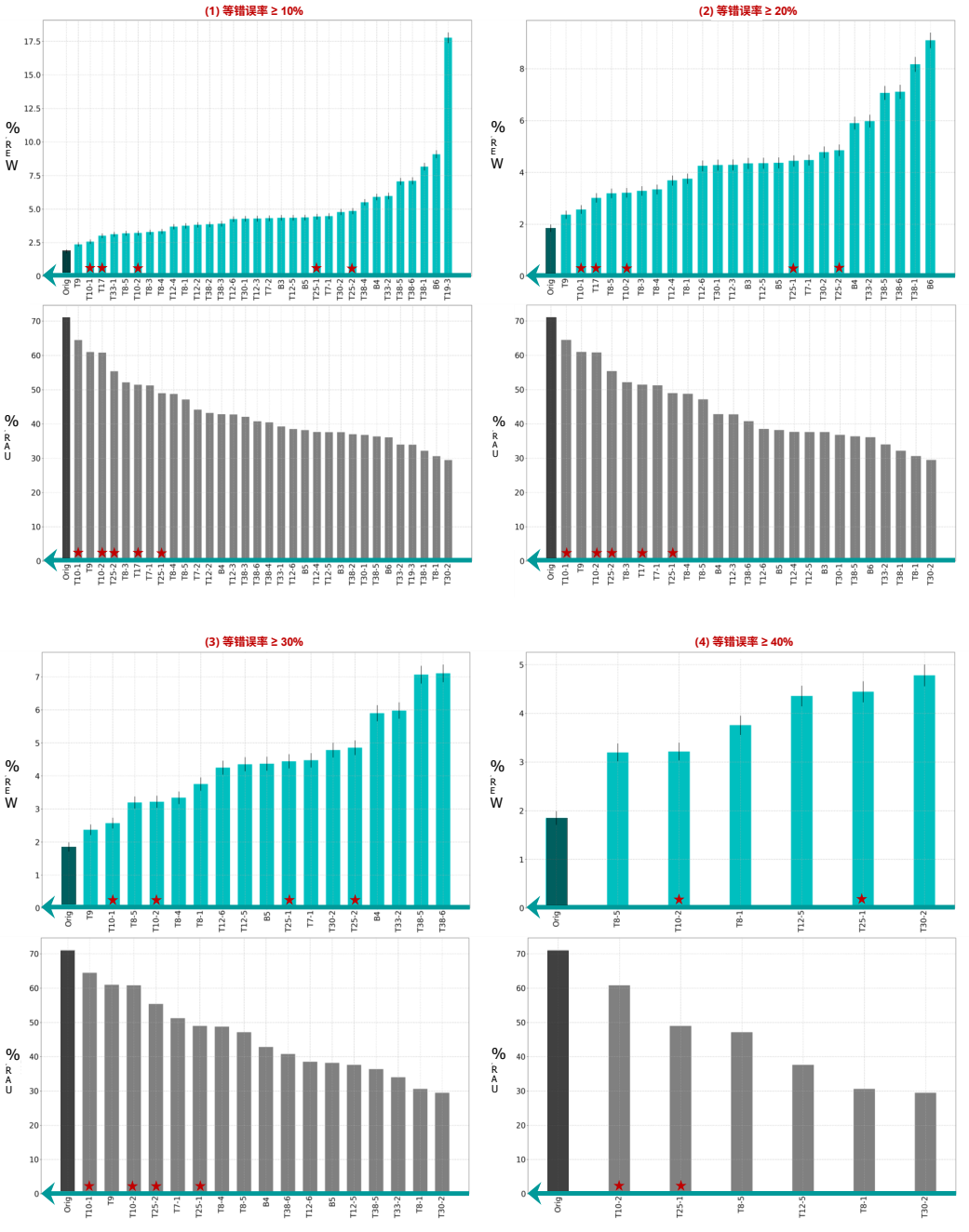


图 8： 根据测试数据中 4 个隐私阈值下的效用指标（词错误率，% 和非加权平均召回率，%）进行的匿名化系统排名（官方排名）。

6.3. 效用指标分析

6.3.1. 语言内容保留（词错误率）

未经处理的测试数据的词错误率为 1.85%。在前三个隐私类别中，内容保留效果最佳的系统**T9**在测试数据上的词错误率为=2.37%，对应于词错误率绝对增加 0.5%、相对增加 28%。在最高隐私类别中，词错误率最低（3.2%）的两个系统——级联系统**T8-5**和基于编解码器的系统**T10-2**——自动语音识别性能下降更为明显：绝对增加约 1.4%，相对增加 73–74%。在最高隐私类别中，词错误率介于 3.2% 至 4.8% 之间。

6.3.2. 情感保留（非加权平均召回率）

原始（未经处理）测试数据上的非加权平均召回率（UAR）在图 7 中以黑色圆圈标示，为 71%。在前三个隐私类别中，表现最佳的系统（**T10-1,2**、**T9**）取得约 61%–65% 的 UAR，而符合规则的系统（**T9**、**T7-1**、**T8-4**）在第三个隐私类别中取得约 49%–61%，在前两个隐私类别中取得 51%–61%（**T9**、**T8-3**、**T7-1**）。在最高隐私类别中，表现最佳的系统（**T10-2**、**T25-1**、**T8-5**）取得约 47%–61% 的 UAR，而最优的符合规则系统（**T8-5**）取得 47%。在最高隐私类别中，符合规则系统中最佳与最差系统（**T8-5** 和 **T30-2**）之间的 UAR 差距超过 17%（该类别所有提交系统之间的差距则为 31%），这突显出显著的性能差异，也凸显了情感感知架构选择的重要性。明确采用情感保留机制的系统（**T9**、**T10-1,2**、**T25-1,2**）在所有隐私类别中的 UAR 排名均高于未采用情感专门设计的系统。

在实现高隐私类别（EER≥30%、40%）的顶尖系统（**T9**、**T10**、**T25-2**）中，也有许多系统在较低隐私类别（EER≥10%、20%）的情感保留方面名列前茅。这种跨隐私阈值的一致性表明，这些系统受益于合理的架构设计，例如多阶段情感感知解缠、情感匹配的伪说话人选择以及混合匿名化方法，而不会为了获得隐私收益而牺牲情感保留。

图 9展示了特定情感的准确率。对于原始语音，四种情感类别的准确率较为均衡。匿名化对不同情感类别的准确率影响各异。尤其是，许多系统无法识别悲伤，这是最难保留且在所有匿名化系统中准确率最低的类别。均衡的系统包括**T10-1**，它能够较好地保留所有情感（每种情感的准确率均为60–70%）。偏向快乐的系统包括**T18-1**、**T14-1**、**2**和**T30-2**，它们的快乐准确率很高（80–94%），但悲伤准确率较低（29–45%）。系统**T8-1**、**T30-1**、**T30-2**、**B3**和**T33-2** 对悲伤的准确率接近0%。

6.4. 隐私-效用权衡

VPC 2024年的结果揭示了所有评估指标之间的根本性权衡，没有任何单一系统能够同时最大化隐私（等错误率）、语言内容保留（词错误率）和情感保留（非加权平均召回率）。挑战赛结果表明，运行空间并非均一，隐私与效用之间的关系会随不同指标和工作点而显著变化。

图 10展示了匿名化系统各项指标相对于原始数据结果的变化散点图（Δ词错误率、 Δ非加权平均召回率和 Δ等错误率）。这些图中的帕累托前沿揭示了不同的权衡模式。第一幅图（隐私与效用（非加权平均召回率））标出了处于前沿的系统，其中 T10-1 实现了最佳平衡（ Δ 非加权平均召回率 ＝ 9%， Δ 等错误率 ＝ 704% ），其次是 T10-2 （ Δ 非加权平均召回率 =14 %， Δ等错误率 ＝ 779%）和 T25-1（Δ非加权平均召回率=31%，Δ等错误率=818%）。相比之下，T8-1的隐私保护最强（Δ等错误率 ＝ 979%），但效用成本也很高（Δ非加权平均召回率=57%）。第二幅图（隐私与效用（词错误率））表明，T9实现了良好的权衡（Δ词错误率= 28%， Δ等错误率 ＝652%），而 T10-1同样表现出色（Δ词错误率 = 39%， Δ等错误率 = 704 %）。第三幅图（双重效用（非加权平均召回率与词错误率））标出了能够同时保留两项效用指标的系统，其中 T9（Δ非加权平均召回率= 14%， Δ词错误率 ＝28%）和 T10-1（Δ非加权平均召回率 ＝ 9%， Δ词错误率 ＝ 39%）是最有效的方法。这些帕累托前沿凸显了系统设计选择对于实现隐私与效用平衡的重要作用。

图 13展示了一个归一化性能矩阵，其中列出了所有系统并按综合得分排序，可直接比较所有提交方案在隐私与效用之间的权衡。挑战赛采用多阈值评估，结果表明，最佳系统的选择取决于应用需求。 混合 系统（**T8-3**、**T8-4**、**T8-5**）表明，按话语动态选择方法可以灵活设定工作点，而混合概率则控制其在隐私-效用谱系上的位置。

6.5. 性别对匿名化的影响

图11展示了测试集中女性和男性说话人子集的等错误率结果，并按差值（等错误率$_{male}$ −等错误率$_{female}$）降序排列。不同系统的基于性别的性能差异差别显著。对未保护数据使用的原始自动说话人验证系统表现出显著的男性偏差（女性说话人的等错误率=8.8%，男性说话人的等错误率=0.4%），这表明男性说话人验证本质上更容易。性别偏差最严重的匿名化系统包括**T17**，其绝对差值为14%（等错误率$_{female}$=34%，等错误率$_{male}$=20%），以及**T33-2**，该系统对女性存在偏差（等错误率$_{female}$=24%，等错误率$_{male}$=37%）。性别均衡的系统包括**B6**、**T8-4**、**T25-1**、**T38-3**和**T33-1**，这些系统将差异维持在0.5%以下，表明其在不同性别间具有一致的匿名化效果。

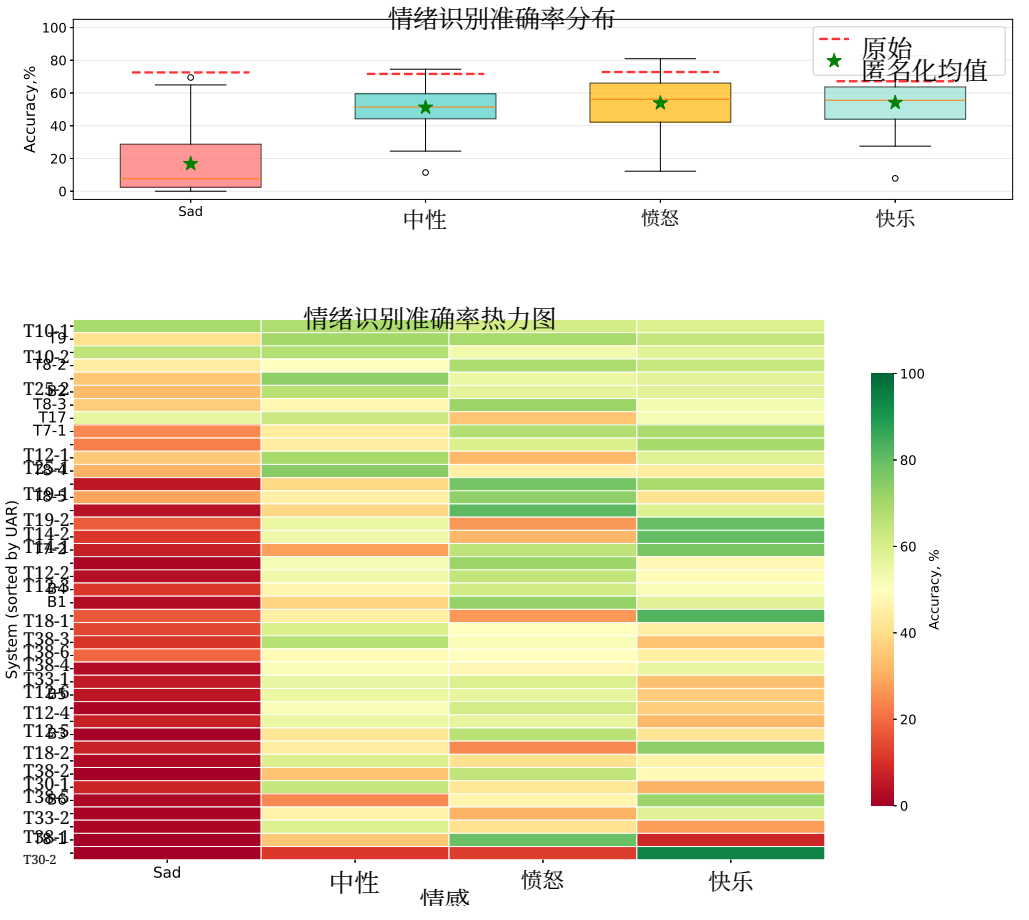


图9： 在*IEMOCAP* 测试集上按情感类别划分的情绪识别准确率

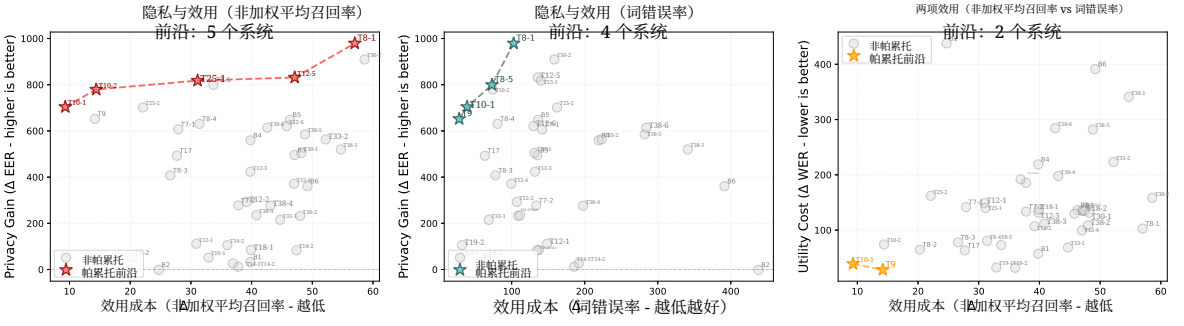


图 10： 测试集上的帕累托曲线分析

为进一步评估匿名化中的性别公平性，我们采用修正公平差异率（mFDR），并在等错误率工作点进行评估，以衡量不同性别之间性能差异的相对幅度。该简化指标针对说话人验证系统和VPC2024评估条件进行了调整，定义如下 [12]：

[12]（Zanon Boito 等人，2022年）采用了类似的方法来评估语音领域中的性别影响。生物识别系统中更一般的公平差异率（FDR）定义可见（Freitas Pereira 和 Marcel，2021年），该指标针对所有人口统计群体采用同一阈值的运行条件进行计算。在 VPC 2024 评估设置中应用此指标（等错误率运行点，两个

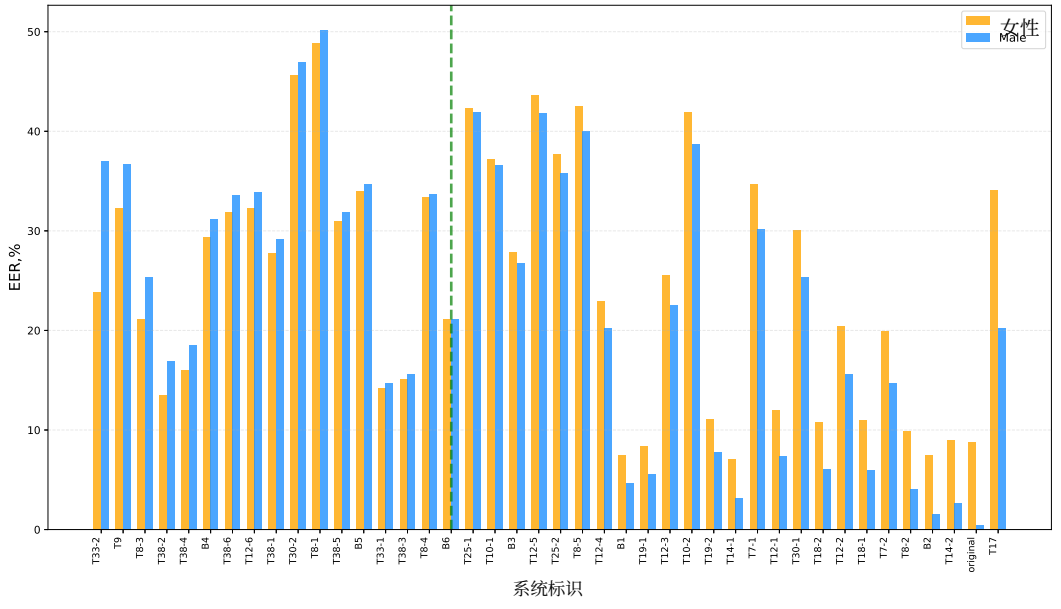


图 11： 匿名化中的性别偏差：按男女等错误率差值排列的系统

$$\text{mFDR} = \frac{2 \cdot |\text{EER}_{\text{female}} - \text{EER}_{\text{male}}|}{(\text{EER}_{\text{female}} + \text{EER}_{\text{male}})} \cdot 100.$$

图12 使用 mFDR 指标，展示了所有匿名化系统在测试数据集和开发数据集上的性别公平性分析。系统根据 mFDR 表现分为不同等级，从 优秀 (mFDR<5%)，适用于 **B6**、**T8-4**、**T25-1**、**T10-1**、**B5**、**T8-1**和 **T38-5**，到 较差，适用于系统 **B2**、**T14-2**、**T8-2**和 **T14-1**。原始数据上的 mFDR 最高，达到 182%。图12 (2)显示开发集和测试集上的 mFDR 具有一致性。对于大多数匿名化系统，mFDR 在测试集上有所改善。

6.6. 基线影响

表 7 展示了基线对提交系统的影响。基线系统（B1–B6）确立了性能界限，参赛系统旨在超越这些界限。B1 引入的基于池选择和平均的 x-vector 匿名化策略，也应用于 T14-1 和 T14-2。多个具有副语言信息保留功能的级联 ASR-TTS 系统基于 B3；该基线采用语音转写和基于 GAN 的伪说话人生成。T12-2、T12-3、T12-4 和 T30-1 直接基于 B3 构建，并引入情感感知和其他匿名化策略。这些系统的性能与 B3 相近，其中 T12-2 和 T12-3 的整体效用略有提升。T12-5 和 T12-6 对 B5（采用基于 VQ 的内容特征的 ASR-BN）进行修改，以原始基频曲线和其移动平均值的加权平均取代原始基频曲线。T12-6 还对所得基频曲线施加白高斯噪声。这些系统的性能接近 B5。相比之下，T25-1 和 T25-2 使用 VQ-BN 提取内容特征，同时引入情感迁移机制，其隐私和 UAR 指标均显著优于 B5。T19-3 也采用了基于 VQ 的方法。神经编解码器基线 B4 的成功推动了社区的广泛采用；T10、T12-1 和 T17 代表了独立的神经编解码器方法，且性能具有竞争力。T10 和 T17 在 WER 和 UAR 两项指标上均显著优于 B4，T10 系统的 EER 值也更高。总体而言，表现最佳的提交系统在所有指标上均显著优于基线。EER 从 34%（B5）提升至 49.5%（T8-1），UAR 从 54%（B2）提升至 65%（T10-1），WER 则从 2.9%（B1）降至 2.37%（T9）。

性别群体），对应于女性和男性说话者在拒真率和误受率绝对差异的加权和。mFDR通过以整体性能水平对差距进行归一化，解决了基于差异的偏差度量在基准错误率较小或不同条件下错误率幅度不同时会失去敏感性的问题。从这个角度看，mFDR的理念更接近于生物特征公平性的基尼聚合率（GARBE）（Howard 等人，2022年；Chouchane 等人，2024年）。

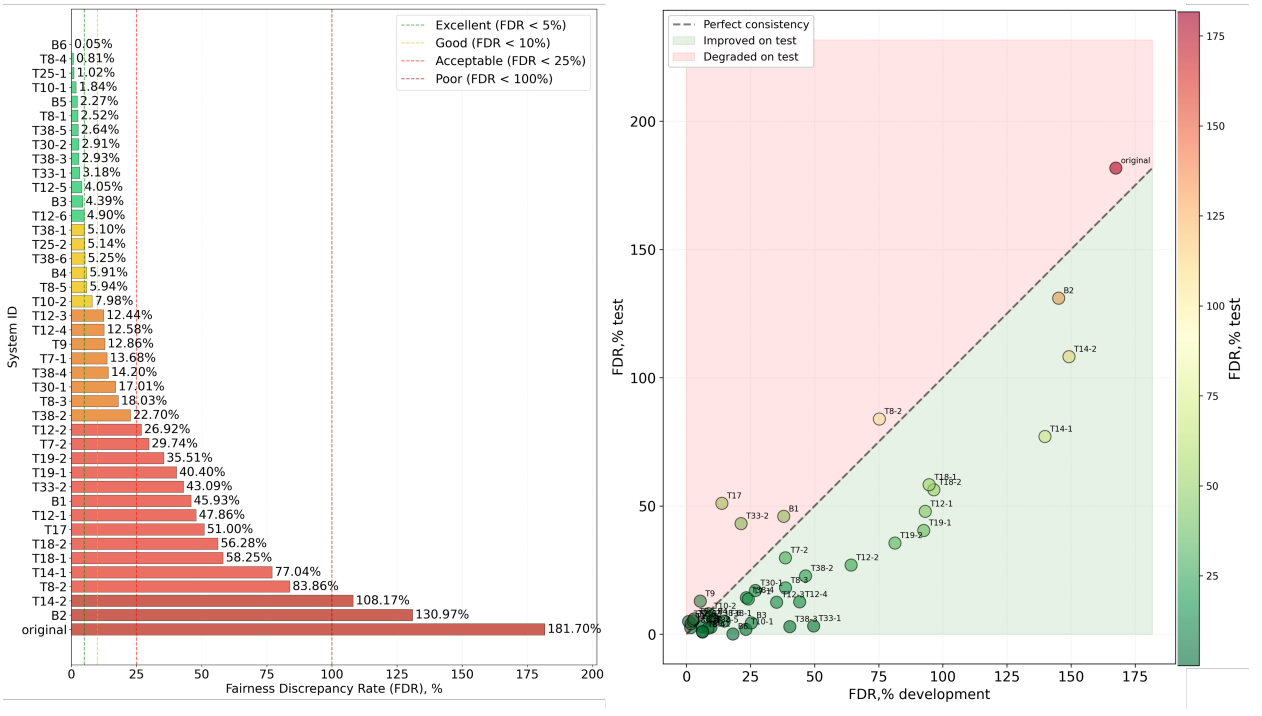


图 12： 匿名化中的性别公平性分析：(1) 根据测试集上的 mFDR 结果进行系统排名；(2) 开发集与测试集之间的 mFDR 一致性。

6.7. 总结

级联方法在绝对隐私方面明显占优（EER = 48%），但无法保留情感。神经编解码器方法实现了最佳的整体平衡，其中 **T10-2** 代表了隐私效用折衷的当前最先进水平。混合方法通过动态方法选择，成功兼顾了不同的优化目标。采用显式属性解耦和自监督学习（SSL）表征的先进语音转换系统，相较于传统语音转换取得了大幅提升，体现了该领域对深度学习进展的应用。挑战赛结果表明，尽管仍能实现强有力的隐私保护，但这需要精心设计架构，以平衡多个相互竞争的目标。从简单的基线方法发展到复杂的多阶段解耦系统，反映了语音匿名化研究中方法的迅速发展。

7. 讨论与未来展望

本节介绍 VPC 2024 年及往届赛事的主要见解，考察评估方法和数据局限性，概述新兴研究方向，并为未来的语音匿名化研究提供战略性建议。

7.1. 评估

可靠的评估是推动匿名化取得进展的关键，同时也有助于确保其在现实世界中的适用性。然而，当前的评估方法存在一些局限，需要仔细分析。

7.1.1. 隐私评估的可靠性

半知情攻击模型（见第2.2节）下的等错误率接近 50%，表明隐私保护达到理想水平。提交至 VPC 的许多系统似乎都实现了如此高水平的隐私保护。然而，近期的若干分析质疑，这些令人鼓舞的匿名化性能估计是否可靠。首届 VoicePrivacy 攻击者挑战赛[13]（Tomashenko 等人，2025a）及其他研究（Panariello 等人，2025年）表明

[13]https://www.voiceprivacychallenge.org/attacker/

等错误率被夸大，可能是攻击者泛化能力较差所致，而非匿名化效果强。与此同时，当前攻击模型存在多项限制，也可能导致隐私保护效果被高估。这些限制包括：

- 数据：攻击者使用各系统生成的匿名化数据训练自动说话人验证系统。在更现实的场景中，攻击者可能会获取任意外部语料库，包括其他匿名化语音数据集，以增强攻击能力。

- 特征：攻击者使用 Fbank 输入特征，并按照典型的自动说话人验证流程训练说话人编码器，而其他研究（Tomashenko 等人，2025c，b；Bakari 等人，2025年）表明，上下文相关时长特征、语音时序动态特征和非音色线索，在匿名化后都能有效揭示说话人特定信息。

- 模型与超参数：攻击采用固定架构（即 ECAPA-TDNN）和未经调优的超参数，而有证据表明，可以使用其他模型实施更强的攻击（Bakari 等人，2025年），例如 *WavLM*。

这些局限性在首届 VoicePrivacy 攻击者挑战赛（Tomashenko 等人、2025a、2024b）中得到了解决。该挑战赛是 ICASSP 2025 SP 大挑战赛的一部分，旨在评估针对若干语音匿名化系统的攻击。采用强攻击对于确保隐私评估的可靠性至关重要。研究考察了七种匿名化系统：VPC 2024 中表现最好的三个基线系统，即 **B3**、**B4** 和**B5** （见第3.4节），以及四个参赛系统。使用基线攻击模型进行评估时，这四个系统（最初）达到了最高隐私水平（等错误率≥40%，图2中的第 4 类），分别为**T8-5**、**T10-2**、**T12-5**和**T25-1**。使用基线攻击者进行评估时，这些系统达到第 4 类隐私水平（等错误率≥40%）。然而，在接受攻击者挑战赛参与者开发的最佳攻击模型测试时，所有系统的等错误率都大幅下降（等错误率≥20%，第 2 类）。结果还表明，使用更强的攻击模型进行评估时，基于等错误率得出的系统排名会发生变化。这些发现凸显了在隐私评估中使用尽可能强的攻击模型，对于确保评估可靠、可信的重要性。

7.1.2. 指标

自 2020 年 VPC 以来，人们提出了种类繁多的评估指标，涵盖隐私和效用两个方面。尽管如此，仍存在若干局限和顾虑：

隐私度量指标。 使用等错误率无法反映法律隐私要求，例如可链接性、个体识别或推断 (第29条工作组，2014)。等错误率可能无法反映再识别风险：即使特定用户或使用场景下成功再识别的概率发生变化，该指标仍可能保持稳定。它只关注单一工作点上的平均表现，因而掩盖了最坏情况下的风险和异常值脆弱性，可能会对隐私保护效果给出过于乐观的评估，尤其是对于匿名化后仍很容易被识别的个人（Nautsch 等人，2020年；Williams 等人，2024年）。

为更好地兼顾技术和法律预期，人们提出了替代性评估框架。Vauquier 等人提出了一种以可链接性和个体识别为基础，并受到法律认可的匿名信息定义解释启发的方法（2025年）。使用这些指标得出的结果表明，匿名化后残留的隐私风险比 EER 估计值所显示的更为显著，这凸显了评估指标需要更好地兼顾技术和法律视角。其他替代指标还包括 Bäckström 等人提出的相似度排名披露（SRD）指标（2025年），其单位为熵（比特）。该指标定义为观察相似度排名后，熵（不确定性）的减少量。它量化了在将查询样本与模板数据库进行比较时，可以从相似度排名中推断出多少关于真实身份的个人可识别信息（PII），并有望成为评估语音匿名化隐私保护效果的指标（Chandra 等人，2025年）。在未来的 VPC 版本中采用能够反映最具挑战性或对隐私最敏感样本表现的指标，或许能给出更符合现实场景预期的匿名化效果估计。

效用指标。当前效用指标的一个局限在于自动指标与感知质量之间存在脱节。虽然更强大的自动语音识别模型通常能获得更低的词错误率，但这并不一定意味着质量有所提升，甚至也不一定表示可懂度有所改善。研究表明，词错误率的估计值并不总是与

主观可懂度（Panariello 等人，2024b），这凸显了自动评估与人类感知之间的根本差距。此外，虽然词错误率的估计值反映了语音内容的保留情况，却忽视了效用的其他重要方面。当前指标未能充分体现语音自然度、韵律保留和声学质量。

建议的解决方案包括开发多维度指标，以评估客观和主观效用；整合针对特定使用场景的评估标准；并设计能够联合评估可懂度、自然度和下游任务效用的效用指标。

7.1.3. 数据与基准测试

此前所有 VPC 版本主要依赖 *LibriSpeech* 数据集来评估隐私和内容保留，2020年版和2022年版也使用了 VCTK。这些语料库包含干净的、单说话人的朗读英语语音。由于后者缺乏情感标注，2024年版引入了 *IEMOCAP* 语料库，该语料库包含自发且富有情感表现力的语音，用于支持情感保留评估。然而，由于说话人数量和规模有限，*IEMOCAP* 本身并不适合用于评估隐私和内容保留。

当前数据集和实验方案存在一些局限性：(1) 隐私与效用评估中的数据集不一致性。 使用一个数据集估计隐私和内容保留，而使用另一个数据集估计情感保留，可能会给结果带来偏差，并引发对系统性能比较是否公平的担忧。(2)语言与口音覆盖范围。 语音数据均为英语，多语言、语码转换或带有浓重口音的语音所占比例极低。(3) 语音特征。*LibriSpeech*仅包含来自人口统计特征有限群体的干净朗读语音；该数据集缺乏自发语音、儿童语音、病理语音，以及能够反映现实世界多样性的说话人差异。(4) 录音条件。 目前尚未针对许多现实应用场景（如电话通话、直播和视频会议）常见的噪声、混响或压缩声学条件，系统评估匿名化鲁棒性。(5) 对话场景。*LibriSpeech* 和 *IEMOCAP* 均为单说话人数据集，而挑战赛中探索的现有解决方案也仅限于单说话人匿名化。这些方法尚不能直接应用于多说话人或语音重叠场景 Miao 等人 (2025a)；Tomashenko 等人(2025d)。

未来的语音匿名化研究需要更大且更多样化的数据集，涵盖从不同年龄、口音、语言、语音风格、情绪状态和声学条件下的说话人收集的数据。为了确保匿名化技术稳健且普遍适用，还需要能够反映更具挑战性和现实性的使用场景的数据集，例如涉及重叠语音和低资源语言的场景；同时还应建立公平性基准，评估匿名化性能在不同说话人群体特征（年龄、性别、口音、语言背景、言语障碍）之间是否保持一致。

7.2. 新兴和先进的匿名化方法

本节综述匿名化领域的最新进展，重点介绍方法创新和特定应用领域的发展。近期大多数研究仍以基于解耦的方法为基础，通过显式地将语音分解为相互独立的因素，实现可控的匿名化。目前已出现几种主要方法：

多属性解耦将单属性方法扩展到多个语音属性的同步解耦（Miao 等人，2025b；Yao 等人，2025年）。这一进展能够实现更精细、考虑上下文的匿名化，并同时控制声音身份、情感表达、人口统计属性及其他说话人特征，为特定领域的应用提供了更大的灵活性。

无说话人嵌入方法（Cai 等，2024年；Franzreb 等，2025年；Tang 等，2025年）完全避免使用显式的说话人嵌入。相反，这些方法直接从语音信号中学习匿名化表征，从而有望简化匿名化流程，并提升在未见过的声学条件和说话人群体上的泛化能力。

近期研究还探讨了其他多样化的应用和具有挑战性的场景。这些研究主要围绕三个维度展开：部署约束、目标人群和范式变体：

语言与跨语言泛化：多语言匿名化方面的新兴研究（Miao 等人，2022b，a；Deng 等人，2023年；Miao 等人，2023年；Meyer 等人，2024年；Yao 等人，2024b）以及语码转换场景（Meyer 等人，2025年）表明，匿名化模型能够泛化到其他语言和混合语言语音。

跨语言泛化扩大了实际应用范围，无需多语言训练数据，也无需重新训练或调整。

对话场景和多说话人录音：许多自然的使用场景都涉及多位说话人同时发言，以及复杂的对话动态，这些因素使匿名化工作更为复杂（Miao 等人，2025a）。近期研究（Tomashenko 等人，2025d）应对了在多说话人重叠对话录音中对单个目标说话人进行匿名化这一挑战，并提出了新方法和改进的评估策略，以增强隐私保护并评估这些复杂场景中的效用。

录音条件和噪声语音：只有少数研究关注在噪声、混响或压缩等声学条件下评估匿名化鲁棒性，而这些条件在许多实际应用中十分常见（例如，电话通话、直播和视频会议）。Miao 等人 （2024年）的研究表明，将匿名化技术应用于来自*VoxCeleb-2*数据集的噪声自发语音，可能会扭曲语音内容。

实时与低延迟匿名化：对于流媒体、视频会议和电话通信等实时应用至关重要。近期研究（Quamer 和 Gutierrez-Osuna，2024年）通过以轻量级卷积神经网络架构和分块语音处理取代计算密集型模型（例如 *HuBERT*、HiFi-GAN），应对了这一挑战，并在严格的延迟限制下实现匿名化，同时避免效用显著下降。

目标人群与专业语音领域：近期研究已将匿名化技术拓展至代表性不足的人群和专业语音领域。针对儿童语音的技术应考虑其独特的声学特征和发育差异（Kulkarni 等人，2025年）。Tayebi Arasteh 等人探索的病理语音（构音障碍、发声障碍等）处理方法，在保护声音身份的同时，保持了语音可懂度和临床诊断效用（ 2024年）。

同步与异步匿名化：目前已出现两种互补的匿名化范式，各自适用于不同的使用场景（Chen 等人，2025年）。同步匿名化会同时改变机器感知和人类感知的声音身份，修改语音信号，使自动系统和人类听者都将其感知为不同的声音身份。这种方法符合 VPC 对语音匿名化的定义。异步匿名化只改变机器可辨识的说话人属性，同时保留人类对原始声音身份的感知。

用户控制与灵活性：近期研究（Hui 等人，2025年）提出了基于大语言模型（LLM）的自动语音识别-文本转语音框架，并利用该框架实现基于提示的语音匿名化。该方法通过自然语言提示指定所需的说话人属性，从而实现灵活、可由用户控制的匿名化，同时通过命名实体识别和替换来保护内容隐私。该框架为依赖上下文和特定任务的匿名化策略开辟了新的可能性。

超越声音身份：多属性隐私保护：除了声音身份隐私，不同场景下的其他属性也可能需要保护。例如，背景环境声音（如火车站的广播）可能会泄露说话人的位置，因此匿名化时除了声音身份，还需要考虑依赖场景的音频线索。内容隐私（Williams 等人，2021年，2022年；Sinha 等人，2024年；Hui 等人，2025年）也是一个主要关注点。尽管如今人们已开始关注后者，大多数方法仍将命名实体视为主要的隐私敏感信息，并直接将其删除——但这种做法并非总是切实可行。即使命名实体被删除，攻击者仍可能从上下文中推断出敏感含义。

可逆性与恢复要求：为确保强有力的隐私保障，大多数匿名化系统都设计为不可逆。然而，在某些使用场景中，可逆性是必要的。例如，在取证或安全场景下，应能够恢复原始内容和说话人身份信息，以便开展调查。近期研究（Zhang 等人，2025年）探索了可恢复匿名化方法，使匿名化语音能够选择性恢复或进行用户定制修改。这一方向旨在满足取证和授权调查场景的需求，在这些场景中，条件解密和语音恢复不可或缺，同时在正常运行时维持隐私。实际实施时，可以允许用户或授权方安全地保存原始与匿名化说话人/内容之间的映射，以便在必要时进行恢复。

8. 结论

VoicePrivacy2024 挑战赛是语音匿名化研究的重要里程碑，展示了隐私保护语音技术取得的显著进展，这些技术能够同时保留语言内容和情感内容。本次挑战赛吸引了来自 16 个国家的 13 个团队提交的 36 份作品，代表

学术机构和非学术机构均积极参与，国际参与度显著提升，方法多样性也大幅增加——参与数量是上一届挑战赛的两倍以上。许多提交的系统已远远超越渐进式的基线改进，引入了结合神经音频编解码器、属性解耦、自监督学习模型和情感感知设计的最先进神经网络架构。

结果表明，强大的隐私保护往往只能通过复杂的神经网络架构，以及对多个相互竞争目标的审慎考量来实现。提交的系统中，只有六个达到了 EER $\geq$ 40% 这一最高隐私目标，而只有四个满足了全部其他挑战赛要求，这说明在保护隐私的同时保留效用存在根本性困难。结果表明，级联自动语音识别+文本转语音系统能够实现近乎完美的匿名化，但代价是情感保留能力下降。通过多阶段解耦和帧级情感蒸馏，基于神经编解码器的方法在隐私与效用要求之间实现了更好的平衡。未显式采用情感感知组件的系统，其表现大多不如配备专门情感蒸馏、编码器集成或情感匹配的伪说话人选择机制的系统。

出现了五个主要技术趋势，包括：采用神经编解码器来提供属性解耦框架；利用自监督学习和预训练模型进行特征提取；对说话人身份、语言内容、韵律进行多阶段属性解耦，并采用情感匿名化技术，通过动态方法选择实现灵活的隐私-效用权衡；以及利用情感编码器、蒸馏机制和情感条件合成进行情感感知设计。

表现最佳的提交系统在所有指标上均显著优于基线。具体而言，各项指标如下：等错误率从 34%（**B5**）提升至 49.5%（**T8-1**），非加权平均召回率从 54%（**B2**）提升至 65%（**T10-1**），词错误率则从 2.9%（**B1**）降至 2.37%（**T9**）。

一些挑战和其他问题仍然存在。其他研究 Tomashenko 等人 (2025a) 指出，隐私评估容易出现高估：四个最初取得等错误率 $\geq$ 40% 的顶尖系统，后来被证明容易遭受攻击，攻击会使等错误率降至不足 20%。这一发现凸显了开展独立且持续的两阶段评估的重要性；在此过程中，由相互竞争的团队开发并改进防御措施与攻击方法。本次挑战赛强调，采用独立或持续的两阶段评估框架至关重要：由不同团队开发匿名化系统和攻击者，确保隐私评估基于经过独立验证的威胁模型，而非单一的标准化攻击者。

新兴研究方向包括多属性和特定应用匿名化、实时且计算高效的解决方案，以及利用基于提示和大语言模型驱动方法的语义级方法，以实现灵活且可由用户控制的匿名化。从简单的基线方法发展到复杂的多阶段解耦系统，反映了语音匿名化研究中方法的快速演进。随着该领域不断发展，并迈向暂定于2026年举行的下一届 VoicePrivacy 挑战赛，解决这些评估局限、拓展应用场景并确保不同人群之间的公平性，对于开发稳健、可信且可广泛部署的语音匿名化解决方案至关重要。

致谢

本研究得到法国国家研究署通过 SpeechPrivacy 项目（https://anr.fr/Projet-ANR-23-CE23-0022）以及由法国网络安全 PEPR 资助的 IPoP 项目（https://www.pepr-cybersecurite.fr/projet/ipop/）的支持。本研究还部分得到日本科学技术振兴机构（JST）PRESTO 资助编号 JPMJPR23P9 的支持。实验使用 Grid' 5000 测试平台进行。挑战赛组织方感谢*Ünal Ege Gaznepoğlu* 在代码库方面提供的帮助。

参考文献

Akti, S., Nguyen, T.N., Liu, Y., Waibel, A., 2024a. 探讨采用不同瓶颈特征的语音转换架构在语音隐私挑战赛 2024 中的应用。VoicePrivacy 2024 挑战赛。Akti, S., Nguyen, T.N., Liu, Y., Waibel, A., 2024b. 语音隐私——探讨采用不同瓶颈特征的语音转换架构，载于：SPSC 2024 论文集。Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G., 2020. Common Voice：一个大规模多语言语音语料库，载于：第十二届语言资源与评测会议（LREC），第 4218–4222 页。Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein 生成对抗网络，载于：国际机器学习会议（ICML），第 214–223 页。

第29条工作组，2014。关于匿名化技术的第 05/2014 号意见。https://ec.europa.eu/justice/article-29/。网址：https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216en.pdf。获欧洲数据保护委员会（EDPB）认可。Baas, M., van Niekerk, B., Kamper, H., 2023. 仅使用最近邻进行语音转换。arXiv 预印本 arXiv:2305.18975。Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., 等，2021。XLS-R：大规模自监督跨语言语音表示学习。arXiv 预印本 arXiv:2111.09296。Bäckström, T., Vali, M.H., Nguyen, M., Rech, S., 2025. 语音与语言处理中相似度排名的隐私泄露。arXiv 预印本 arXiv:2508.05250。Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. wav2vec 2.0：一种用于自监督语音表示学习的框架。《神经信息处理系统进展》33，12449–12460。Bakari, R., Blouch, O.L., Evans, N., Gengembre, N., Panariello, M., Todisco, M., 2025. 非音色线索对语音匿名化及其评估的影响，载于：第五届语音通信安全与隐私研讨会，第 35–42 页。doi：10.21437/SPSC.2025-6。Bogdanov, D., Won, M., Tovstogan, P., Porter, A., Serra, X., 2019. 用于自动音乐标注的 MTG-Jamendo 数据集，载于：ICML 2019 音乐发现机器学习研讨会。Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., Zeghidour, N., 2023. AudioLM：一种音频生成的语言建模方法。arXiv 预印本 arXiv:2209.03143。Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. 德语情感语音数据库，载于：Interspeech，第 1517–1520 页。Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. IEMOCAP：情感双人动作捕捉数据库。《语言资源与评测期刊》42，335–359。Cai, Z., Xinyuan, H.L., Garg, A., García-Perera, L.P., Duh, K., Khudanpur, S., Andrews, N., Wiesner, M., 2024. 情感保留型说话人匿名化中的隐私与情感保留权衡，载于：SLT 论文集，IEEE，第 409–414 页。Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014. CREMA-D：众包情感多模态演员数据集。《IEEE 情感计算汇刊》5，377–390。Champion, P., 2023. 语音匿名化：评估与设计说话人匿名化技术。博士论文。洛林大学。Chandra, S., Petteno, M., Evans, N., Bäckström, T., Kolossa, D., Martin, R., 2025. 使用相似度排名披露评估语音匿名化，载于：编写中。Chen, L., Guo, C., Wang, R., Lee, K.A., Ling, Z.H., 2025. 用于异步语音匿名化的任意到任意说话人属性扰动。《IEEE 信息取证与安全汇刊》。Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Wei, F., 2022. WavLM：面向全栈语音处理的大规模自监督预训练。《IEEE 信号处理选题期刊》16，1505–1518。Cho, K., van Merriënboer, B., Bahdanau, D., Bengio, Y., 2014. 神经机器翻译的特性：编码器—解码器方法，载于：Wu, D., Carpuat, M., Carreras, X., Vecchi, E.M.（编），《SSST-8 论文集：第八届统计机器翻译语法、语义与结构研讨会》，计算语言学协会，多哈，卡塔尔，第 103–111 页。网址：https://aclanthology.org/W14-4012/，doi：10.3115/v1/W14-4012。Chouchane, O., Busch, C., Galdi, C., Evans, N., Todisco, M., 2024. 自动说话人验证公平性评估中差异化性能指标的比较，载于：Odyssey 2024 论文集，第 209–216 页。Chung, J.S., Nagrani, A., Zisserman, A., 2018. VoxCeleb2：深度说话人识别，载于：Interspeech，第 1086–1090 页。Chung, Y.A., Zhang, Y., Han, W., Chiu, C.C., Qin, J., Pang, R., Wu, Y., 2021. W2v-BERT：结合对比学习与掩码语言建模的自监督语音预训练，载于：2021 年 IEEE 自动语音识别与理解研讨会（ASRU），IEEE，第 244–250 页。Das, A., Franzreb, C., Herzig, T., Pirlet, P., Polzehl, T., 2024. VoicePrivacy 挑战赛 2024 系统描述，载于：VoicePrivacy 2024 挑战赛。Défossez, A., Copet, J., Synnaeve, G., Adi, Y., 2023. 高保真神经音频压缩。《机器学习研究汇刊》网址：https://openreview.net/forum?id=ivCd8z8zR2。Deng, J., Teng, F., Chen, Y., Chen, X., Wang, Z., Xu, W., 2023。 {V-Cloak}：兼顾可懂度、自然度与 {音色保留}{实时} 语音匿名化，载于：第 32 届 USENIX 安全研讨会（USENIX Security 23），第 5181–5198 页。Desplanques, B., Thienpondt, J., Demuynck, K., 2020. ECAPA-TDNN：用于基于 TDNN 的说话人验证的强化通道注意力、传播与聚合，载于：Interspeech，第 3830–3834 页。Dubey, H., Gopal, V., Cutler, R., Aazami, A., Matusevych, S., Braun, S., Eskimez, S.E., Thakker, M., Yoshioka, T., Gamper, H., Aichner, R., 2022. ICASSP 2022 深度噪声抑制挑战赛，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），第 9271–9275 页。Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE：慕尼黑多功能快速开源音频特征提取器，载于：第 18 届 ACM 多媒体国际会议论文集，计算机协会，美国纽约州纽约市，第 1459–1462 页。网址：https://doi.org/10.1145/1873951.1874246，doi：10.1145/1873951.1874246。Fonseca, E., Favory, X., Pons, J., Font, F., Serra, X., 2022. FSD50K：一个人工标注声音事件的开放数据集。《IEEE/ACM 音频、语音与语言处理汇刊》30，829–852。doi：10.1109/TASLP.2021.3133208。Franzreb, C., Das, A., Polzehl, T., Möller, S., 2025. Private knn-vc：转换语音的可解释匿名化，载于：Interspeech 论文集。Freitas Pereira, T., Marcel, S., 2021. 生物特征领域的公平性：评估生物特征验证系统的一项衡量指标。《IEEE 生物识别、行为与身份科学汇刊》4，19–29。Gemmeke, J.F., Ellis, D.P.W., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M., 2017. Audio Set：用于音频事件的本体与人工标注数据集，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），第 776–780 页。doi：10.1109/ICASSP.2017.7952261。

Ghosh, S., Tyagi, U., Ramaneswaran, S., Srivastava, H., Manocha, D., 2023年. Mmer：用于语音情感识别的多模态多任务学习，载于：Interspeech 2023，页码 1209–1213。doi：10.21437/Interspeech.2023-2271。Gu, W., Liu, Z., Chen, L., Wang, R., Guo, C., Guo, W., Lee, K.A., Ling, Z.H., 2024年。USTC-PolyU 系统：参加 VoicePrivacy 2024 挑战赛。Haq, S., Jackson, P.J., Edge, J., 2009年。说话人相关的视听情感识别，载于：国际视听语音处理会议（AVSP），页码 53–58。He, J., Zhou, J., Sun, H., Wang, H., Qin, Y., 2024年。经韵律匿名化增强的基于 FaCodec 的匿名化方案。2024 年 VoicePrivacy 挑战赛。Howard, J.J., Laird, E.J., Rubin, R.E., Sirotin, Y.B., Tipton, J.L., Vemury, A.R., 2022年。评估针对人脸识别算法提出的公平性模型，载于：国际模式识别会议，Springer，页码 431–447。Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021年。HuBERT：通过掩码预测隐藏单元进行自监督语音表征学习。《IEEE/ACM 音频、语音与语言处理汇刊》29，3451–3460。Hua, H., Shang, Z., Li, X., Shi, P., Yang, C., Wang, L., Zhang, P., 2024年。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：SPSC 2024 论文集，页码 55–60。Hui, B.S.H., Miao, X., Wang, X., 2025年。Securespeech：基于提示的说话人与内容保护。arXiv 预印本 arXiv:2507.07799。Ito, K., Johnson, L., 2017年。LJ 语音数据集。https://keithito.com/LJ-Speech-Dataset/。Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., Wu, Z., Qin, T., Li, X.Y., Ye, W., Zhang, S., Bian, J., He, L., Li, J., Zhao, S., 2024年。NaturalSpeech 3：使用因子化编解码器和扩散模型进行零样本语音合成。arXiv 预印本 arXiv:2403.03100。Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., Likhomanenko, T., Synnaeve, G., Joulin, A., Mohamed, A., Dupoux, E., 2020年。Libri-light：一个用于有限监督或无监督自动语音识别的基准，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），页码 7669–7673。Kasi, K., Zahorian, S.A., 2002年。又一种基频跟踪算法，载于：2002 IEEE 国际声学、语音与信号处理会议，页码 I–361–I–364。doi：10.1109/ICASSP.2002.5743729。Kim, J., Kong, J., Son, J., 2021年。采用对抗学习的条件变分自编码器实现端到端文本转语音，载于：国际机器学习会议（ICML），页码 5530–5540。Ko, T., Peddinti, V., Povey, D., Seltzer, M.L., Khudanpur, S., 2017年。用于鲁棒语音识别的混响语音数据增强研究，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），页码 5220–5224。Kong, J., Kim, J., Bae, J., 2020年。HiFi-GAN：用于高效高保真语音合成的生成对抗网络。《神经信息处理系统进展》33，17022–17033。Kulkarni, A., Teixeira, F., Hermann, E., Rolland, T., Trancoso, I., Doss, M.M., 2025年。儿童语音隐私：初步探索与新兴挑战，载于：Interspeech 2025，页码 2810–2814。doi：10.21437/Interspeech.2025-2117。Kuzmin, N., Luong, H.T., Yao, J., Xie, L., Lee, K.A., Chng, E.S., 2024年。参加 VoicePrivacy 2024 挑战赛的 NTU-NPU 系统。arXiv 预印本 arXiv:2410.02371。Le Blouch, O., Bakari, R., Gengembre, N., 2024a。参加 2024 年 VoicePrivacy 挑战赛的 Orange SHIVA 系统说明。2024 年 VoicePrivacy 挑战赛。Le Blouch, O., Bakari, R., Gengembre, N., 2024b。为 2024 年语音隐私挑战赛调优 DISSC，载于：SPSC 2024 论文集。Leang, S., Augusma, A., Vaufreydaz, D., Castelli, E., Sam, S., Letué, F., 2024年。探索在说话人匿名化中使用带有韵律参数的向量量化变分自编码器。2024 年 VoicePrivacy 挑战赛。Lee, J., Park, T., You, Y., 2024a。2024 年语音隐私挑战赛系统说明。2024 年 VoicePrivacy 挑战赛。Lee, J., Park, T., You, Y., 2024b。利用融合情感增强特征的自动语音识别和文本转语音模型进行语音匿名化，载于：SPSC 2024 论文集，页码 50–54。Li, J., Tu, W., Xiao, L., 2023年。FreeVC：迈向高质量无文本单次语音转换，载于：ICASSP 论文集，IEEE，页码 1–5。Li, Y., Zheng, Y., Fang, J., Chen, J., 2024年。使用 FreeVC 框架进行情感增强说话人匿名化。2024 年 VoicePrivacy 挑战赛。Livingstone, S.R., Russo, F.A., 2018年。瑞尔森情感语音与歌曲视听数据库（RAVDESS）：北美英语中面部与声音表达的动态多模态集合。《PLoS ONE》13，e0196391。Lotfian, R., Busso, C., 2019年。通过从现有播客录音中提取情感语音，构建自然且情感均衡的语音语料库。《IEEE 情感计算汇刊》10，471–483。doi：10.1109/TAFFC.2017.2736999。Lux, F., Koch, J., Schweitzer, A., Vu, N.T., 2021年。参加 Blizzard Challenge 2021 的 IMS Toucan 系统，载于：Blizzard Challenge 研讨会。Maimon, G., Adi, Y., 2023a。使用离散自监督单元在波形域中进行说话风格转换，载于：Bouamor, H., Pino, J., Bali, K.（编），计算语言学协会年会论文集：EMNLP 2023，计算语言学协会，新加坡，页码 8048–8061。网址：https://aclanthology.org/2023.findings-emnlp.541/，doi：10.18653/v1/2023.findings-emnlp.541。Maimon, G., Adi, Y., 2023b。使用离散自监督单元在波形域中进行说话风格转换。网址：https://arxiv.org/ abs/2212.09730，arXiv:2212.09730。McAdams, S., 1984。频谱融合、频谱解析与听觉意象的形成。博士论文。斯坦福大学。Meyer, S., Kolos, E., Vu, N.T., 2025年。迈向语码转换语音匿名化的初步探索。Interspeech 论文集。Meyer, S., Lux, F., Koch, J., Denisov, P., Tilli, P., Vu, N.T., 2023a。韵律并非身份：一种使用韵律克隆的说话人匿名化方法，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），页码 1–5。Meyer, S., Lux, F., Vu, N.T., 2024年。探究多语种说话人匿名化的可行性。已被 INTERSPEECH 2024 接收。Meyer, S., Miao, X., Vu, N.T., 2023b。VoicePAT：用于语音隐私研究的高效开源评估工具包。《IEEE 开放获取信号处理期刊》。Meyer, S., Tilli, P., Denisov, P., Lux, F., Koch, J., Vu, N.T., 2023c。使用生成对抗网络对语音进行匿名化以保护说话人隐私，载于：IEEE 口语语言技术研讨会（SLT），页码 912–919。

第三届 VoicePrivacy 挑战赛：在语音匿名化中保留情感表现力和语言内容Miao, X., Tao, R., Zeng, C., Wang, X., 2025a。多说话人匿名化基准。《IEEE 信息取证与安全汇刊》20，3819–3833。Miao, X., Wang, X., Cooper, E., Yamagishi, J., Evans, N., Todisco, M., Bonastre, J.F., Rouvier, M., 2024年。Synvox2：迈向隐私友好的 VoxCeleb2 数据集，载于：ICASSP 2024-2024 IEEE 国际声学、语音与信号处理会议（ICASSP），IEEE，页码 11421–11425。Miao, X., Wang, X., Cooper, E., Yamagishi, J., Tomashenko, N., 2022a。在未见条件下分析语言无关的说话人匿名化框架，载于：Interspeech 2022 论文集，页码 4426–4430。Miao, X., Wang, X., Cooper, E., Yamagishi, J., Tomashenko, N., 2022b。使用自监督预训练模型的语言无关说话人匿名化方法。arXiv 预印本 arXiv:2202.13097。Miao, X., Wang, X., Cooper, E., Yamagishi, J., Tomashenko, N., 2023年。使用正交 Householder 神经网络进行说话人匿名化。《IEEE/ACM 音频、语音与语言处理汇刊》31，3681–3695。doi：10.1109/TASLP.2023.3313429。Miao, X., Zhang, Y., Wang, X., Tomashenko, N., Soh, D.C.L., Mcloughlin, I., 2025b。调整通用的基于解缠的说话人匿名化方法，以增强情感保留。《计算机语音与语言》，101810。Nautsch, A., Jimenez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M.A., Mtibaa, A., Abdelraheem, M.A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., Bonastre, J.F., Busch, C., 2019年。保护说话人和语音特征描述中的隐私。《计算机语音与语言》58，441–480。Nautsch, A., Patino, J., Tomashenko, N., Yamagishi, J., Noé, P.G., Bonastre, J.F., Todisco, M., Evans, N., 2020年。隐私 ZEBRA：零证据生物特征识别评估，载于：Interspeech，页码 1698–1702。Noé, P.G., Bonastre, J.F., Matrouf, D., Tomashenko, N., Nautsch, A., Evans, N., 2020年。使用语音相似度矩阵评估语音假名化，载于：Interspeech，页码 1718–1722。Noé, P.G., Nautsch, A., Evans, N., Patino, J., Bonastre, J.F., Tomashenko, N., Matrouf, D., 2022年。迈向语音假名化的统一评估框架。《计算机语音与语言》72，101299。O’Brien, B., Tomashenko, N., Chanclu, A., Bonastre, J.F., 2021年。匿名说话人簇：通过聚类界面区分匿名化语音录音，载于：Interspeech，页码 3580–3584。Panariello, M., Meyer, S., Champion, P., Miao, X., Todisco, M., Vu, N.T., Evans, N., 2025年。语音匿名化中过度估计隐私保护的风险与检测。SPSC 2025。Panariello, M., Nespoli, F., Todisco, M., Evans, N., 2024a。使用神经音频编解码器语言模型进行说话人匿名化，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），页码 4725–4729。Panariello, M., Tomashenko, N., Wang, X., Miao, X., Champion, P., Nourtel, H., Todisco, M., Evans, N., Vincent, E., Yamagishi, J., 2024b。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》32，3477–3491。doi：10.1109/TASLP.2024.3430530。Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015年。LibriSpeech：基于公有领域有声读物的自动语音识别语料库，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），页码 5206–5210。Pappagari, R., Wang, T., Villalba, J., Chen, N., Dehak, N., 2020年。X 向量与情感相遇：情感与说话人识别之间依赖关系的研究，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），页码 7169–7173。Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., Evans, N., 2021年。使用 McAdams 系数进行说话人匿名化，载于：Interspeech，页码 1099–1103。Peng, Y., Dalmia, S., Lane, I., Watanabe, S., 2022年。Branchformer：并行 MLP-注意力架构，通过捕获局部和全局上下文实现语音识别与理解，载于：国际机器学习会议（ICML），页码 17627–17643。Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., Khudanpur, S., 2018年。用于深度神经网络的半正交低秩矩阵分解，载于：Interspeech，页码 3743–3747。Qian, J., Han, F., Hou, J., Zhang, C., Wang, Y., Li, X.Y., 2018年。迈向保护隐私的语音数据发布，载于：IEEE 计算机通信会议（INFOCOM），页码 1079–1087。Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.I., Cox, D., Hasegawa-Johnson, M., Chang, S., 2022年。ContentVec：通过解缠说话人改进自监督语音表征，载于：国际机器学习会议，PMLR，页码 18003–18017。Quamer, W., Gutierrez-Osuna, R., 2024年。用于低延迟语音匿名化的端到端流式模型，载于：2024 IEEE 口语语言技术研讨会（SLT），IEEE，页码 727–734。Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023年。通过大规模弱监督实现鲁棒语音识别，载于：国际机器学习会议，PMLR，页码 28492–28518。Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.C., Yeh, S.L., Fu, S.W., Liao, C.F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R.D., Bengio, Y., 2021年。SpeechBrain：通用语音工具包。arXiv 预印本 arXiv:2106.04624。Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.Y., 2020年。FastSpeech 2：快速、高质量的端到端文本转语音，载于：国际学习表征会议（ICLR）。Sharma, G., Dhall, A., Cai, J., 2021年。视听自动群体情感分析。《IEEE 情感计算汇刊》14，1056–1069。Sinha, Y., Raivakhovskyi, M., Schubert, M., Siegert, I., 2024年。保护语音内容风格：超越说话人身份增强隐私。SPSC 2024 论文集，92–101。Snyder, D., Chen, G., Povey, D., 2015年。MUSAN：音乐、语音和噪声语料库。arXiv:1510.08484。Srivastava, B.M.L., Maouche, M., Sahidullah, M., Vincent, E., Bellet, A., Tommasi, M., Tomashenko, N., Wang, X., Yamagishi, J., 2022年。基于 x 向量的说话人匿名化中的隐私与效用。《IEEE/ACM 音频、语音与语言处理汇刊》30，2383–2395。Srivastava, B.M.L., Tomashenko, N., Wang, X., Vincent, E., Yamagishi, J., Maouche, M., Bellet, A., Tommasi, M., 2020a。基于 x 向量的说话人匿名化设计选择，载于：Interspeech，页码 1713–1717。

Srivastava, B.M.L.、Vauquier, N.、Sahidullah, M.、Bellet, A.、Tommasi, M.、Vincent, E.，2020b。在 IEEE 国际声学、语音与信号处理会议（ICASSP）中评估基于语音转换的隐私保护对抗知情攻击者的效果，第 2802–2806 页。Tan, T.、Liu, S.、Duan, Y.、Zhao, S.、Shao, X.，2024。系统描述：采用情感迁移和特征插值的说话人匿名化系统。VoicePrivacy 2024 挑战赛。Tang, B.、Miao, X.、Wang, X.、Li, M.，2025。Sef-mk：通过多重 k 均值量化实现无说话人嵌入语音匿名化。ASRU 论文集。Tayebi Arasteh, S.、Arias-Vergara, T.、Pérez-Toro, P.A.、Weise, T.、Packhäuser, K.、Schuster, M.、Noeth, E.、Maier, A.、Yang, S.H.，2024。应对说话人匿名化中的挑战，在确保病理语音隐私的同时维持效用。Communications Medicine 4，182。Thienpondt, J.、Demuynck, K.，2023。ECAPA2：用于稳健说话人嵌入的混合神经网络架构与训练策略，在：IEEE 自动语音识别与理解研讨会（ASRU），第 1–8 页。Tomashenko, N.、Miao, X.、Champion, P.、Meyer, S.、Wang, X.、Vincent, E.、Panariello, M.、Evans, N.、Yamagishi, J.、Todisco, M.，2024a。VoicePrivacy 2024 挑战赛评估计划。arXiv 预印本 arXiv:2404.02677。Tomashenko, N.、Miao, X.、Vincent, E.、Yamagishi, J.，2024b。首届 VoicePrivacy 攻击者挑战赛评估计划。arXiv 预印本 arXiv:2410.07428。Tomashenko, N.、Miao, X.、Vincent, E.、Yamagishi, J.，2025a。首届 VoicePrivacy 攻击者挑战赛，在：ICASSP 2025—2025 IEEE 国际声学、语音与信号处理会议（ICASSP），IEEE，第 1–2 页。Tomashenko, N.、Srivastava, B.M.L.、Wang, X.、Vincent, E.、Nautsch, A.、Yamagishi, J.、Evans, N.、Patino, J.、Bonastre, J.F.、Noé, P.G.、Todisco, M.，2020a。VoicePrivacy 2020 挑战赛评估计划。
https://www.voiceprivacychallenge.org/vp2020/docs/VoicePrivacy_2020_ _ _ 1_4 _ 评估计划 v .pdf。Tomashenko, N.、Srivastava, B.M.L.、Wang, X.、Vincent, E.、Nautsch, A.、Yamagishi, J.、Evans, N.、Patino, J.、Bonastre, J.F.、Noé, P.G.、Todisco, M.，2020b。介绍 VoicePrivacy 倡议，在：Interspeech，第 1693–1697 页。doi: 10.21437/Interspeech.2020-1333。Tomashenko, N.、Vincent, E.、Tommasi, M.，2025b。在说话人验证和语音匿名化背景下分析语音时间动态，在：ICASSP 2025—2025 IEEE 国际声学、语音与信号处理会议（ICASSP），IEEE，第 1–5 页。Tomashenko, N.、Vincent, E.、Tommasi, M.，2025c。利用上下文相关时长特征构建语音匿名化攻击系统，在：Interspeech 2025。Tomashenko, N.、Wang, X.、Miao, X.、Nourtel, H.、Champion, P.、Todisco, M.、Vincent, E.、Evans, N.、Yamagishi, J.、Bonastre, J.F.，2022a。VoicePrivacy 2022 挑战赛评估计划。arXiv 预印本 arXiv:2203.12468。Tomashenko, N.、Wang, X.、Miao, X.、Nourtel, H.、Champion, P.、Todisco, M.、Vincent, E.、Evans, N.、Yamagishi, J.、Bonastre, J.F.、Panariello, M.，2022b。VoicePrivacy 2022 挑战赛 URL: https://www.voiceprivacychallenge.org/vp2022/docs/VoicePrivacy_2022_ 挑战赛___Natalia_Tomashenko.pdf。Tomashenko, N.、Wang, X.、Vincent, E.、Patino, J.、Srivastava, B.M.L.、Noé, P.G.、Nautsch, A.、Evans, N.、Yamagishi, J.、O'Brien, B.、Chanclu, A.、Bonastre, J.F.、Todisco, M.、Maouche, M.，2021。论文补充材料。《VoicePrivacy 2020 挑战赛：结果与发现》。
https://hal.archives-ouvertes.fr/hal-03335126。Tomashenko, N.、Wang, X.、Vincent, E.、Patino, J.、Srivastava, B.M.L.、Noé, P.G.、Nautsch, A.、Evans, N.、Yamagishi, J.、O'Brien, B.、Chanclu, A.、Bonastre, J.F.、Todisco, M.、Maouche, M.，2022 74, 101362 2020101016202210136 2c。《VoicePrivacy 挑战赛：结果与发现》。Computer Speech and Language。doi：https://doi.org/ 。/ j.csl. 。。Tomashenko, N.、Yamagishi, J.、Wang, X.、Liu, Y.、Vincent, E.，2025d。多说话人录音中的目标说话人匿名化。arXiv 预印本 arXiv:2510.09307。Vauquier, N.、Srivastava, B.M.L.、Hosseini, S.A.、Vincent, E.，2025。语音匿名化的经法律验证的评估框架，在：Interspeech 2025，第 3229–3233 页。doi：10.21437/Interspeech.2025-1699。Veaux, C.、Yamagishi, J.、MacDonald, K.，2019。CSTR VCTK 语料库：用于 CSTR 语音克隆工具包的英语多说话人语料库（版本 0.92）。
https://datashare.is.ed.ac.uk/handle/10283/3443。Wagner, J.、Triantafyllopoulos, A.、Wierstorf, H.、Schmitt, M.、Burkhardt, F.、Eyben, F.、Schuller, B.W.，2023。语音情感识别的 Transformer 时代曙光：弥合效价差距。IEEE 模式分析与机器智能汇刊。Wang, C.、Chen, S.、Wu, Y.、Zhang, Z.、Zhou, L.、Liu, S.、Chen, Z.、Liu, Y.、Wang, H.、Li, J.、He, L.、Zhao, S.、Wei, F.，2023。神经编解码器语言模型是零样本文本转语音合成器。arXiv 预印本 arXiv:2301.02111。Wang, D.、Deng, L.、Yeung, Y.T.、Chen, X.、Liu, X.、Meng, H.，2021。Vqmivc：基于向量量化和互信息的无监督语音表征解耦，用于单样本语音转换，在：Interspeech 论文集，第 1344–1348 页。Wang, Y.、Stanton, D.、Zhang, Y.、Ryan, R.S.、Battenberg, E.、Shor, J.、Xiao, Y.、Ren, F.、Jia, Y.、Saurous, R.A.，2018。风格令牌：端到端语音合成中的无监督风格建模、控制与迁移，在：国际机器学习会议（ICML），第 5180–5189 页。Williams, J.、Pizzi, K.、Das, S.、Noé, P.G.，2022。语音与音频内容隐私的新挑战，在：SPSC 2022 论文集，第 1–6 页。Williams, J.、Pizzi, K.、Tomashenko, N.、Das, S.，2024。匿名化说话人声音：易于模仿，却难以识别？，在：ICASSP 2024—2024 IEEE 国际声学、语音与信号处理会议（ICASSP），IEEE，第 12491–12495 页。Williams, J.、Yamagishi, J.、Valentini-Botinhao, C.、Bonastre, J.F.等，2021。重新审视语音内容隐私，在：SPSC 2021 论文集，第 42–46 页。Xinyuan, H.L.、Cai, Z.、Garg, A.、Duh, K.、García-Perera, L.P.、Khudanpur, S.、Andrews, N.、Wiesner, M.，2024。HLTCOE JHU 向 VoicePrivacy 2024 挑战赛提交的系统。arXiv 预印本 arXiv:2409.08913。Yao, J.、Kuzmin, N.、Wang, Q.、Guo, P.、Ning, Z.、Guo, D.、Lee, K.A.、Chng, E.S.、Xie, L.，2024a。NPU-NTU 系统参加 VoicePrivacy 2024 挑战赛。arXiv 预印本 arXiv:2409.04173。Yao, J.、Liu, H.、Chng, E.S.、Xie, L.，2025。EASY：通过因子化蒸馏实现情感感知说话人匿名化，在：Interspeech 2025，第 3219–3223 页。doi：10.21437/Interspeech.2025-194。Yao, J.、Wang, Q.、Guo, P.、Ning, Z.、Yang, Y.、Pan, Y.、Xie, L.，2024b。MUSA：通过串行解耦实现多语言说话人匿名化。arXiv 预印本 arXiv:2407.11629。

Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.P., 2018年。真实场景下的多模态语言分析：CMU-MOSEI数据集与可解释动态融合图，载于：第56届 ACL 年会（第1卷：长篇论文），第2236–2246页。Zanon Boito, M., Besacier, L., Tomashenko, N., Estève, Y., 2022年。自监督模型中的性别对语音转文本系统的影响研究，载于：Interspeech 2022 论文集，第1278–1282页。Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R.J., Jia, Y., Chen, Z., Wu, Y., 2019年。LibriTTS：一个从 LibriSpeech 派生的文本转语音语料库，载于：Interspeech，第1526–1530页。Zhang, J., Liu, S., Hou, J., Wang, Z., Yu, H., Li, X.Y., 2025年。 {SpeechGuard}：可恢复且可定制的语音隐私保护，载于：第34届 USENIX安全研讨会（USENIX Security 25），第5931–5948页。Zhou, K., Sisman, B., Liu, R., Li, H., 2021年。基于新情感语音数据集的语音转换已见与未见情感风格迁移，载于：IEEE 国际声学、语音与信号处理会议（ICASSP），第920–924页。

A. 训练资源

表 8列出了匿名化系统允许使用的训练资源。

表 8：用于训练匿名化系统的模型和数据列表。 ⋆ 表示不同团队针对相应模型/数据集提出的多个请求。

#	模型
1	WavLM基础版和大型版（Chen 等人，2022年） ⋆ ⋆ ⋆⋆ https://github.com/microsoft/unilm/tree/master/wavlm
2	Whisper（Radford 等人，2023年） ⋆⋆ https://github.com/openai/whisper
3	HuBERT（Hsu 等人，2021年） ⋆ ⋆ ⋆ https://github.com/facebookresearch/fairseq/blob/main/examples/hubert
4	XLS-R（Babu 等人，2021年） ⋆⋆ https://github.com/facebookresearch/fairseq/blob/main/examples/wav2vec/xlsr
5	wav2vec 2.0（Baevski 等人，2020年） ⋆ ⋆ ⋆ https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec https://dl.fbaipublicfiles.com/voxpopuli/models/wav2vec2_large_west_germanic_v2.pt
6	wav2vec2-large-robust-12-ft-emotion-msp-dim(Wagner 等人，2023年) https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim
7	ContentVec (Qian 等人，2022年); https://github.com/auspicious3000/contentvec
8	w2v-BERT(Chung 等人，2021年) https://github.com/facebookresearch/fairseq/tree/ust/examples/w2vbert
9	ECAPA2（Thienpondt 和 Demuynck，2023年）； https://huggingface.co/Jenthe/ECAPA2
10	ECAPA-TDNN（Desplanques 等人，2020年）； https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb
11	NaturalSpeech 3（Ju 等人，2024年）； https://huggingface.co/amphion/naturalspeech3_facodec
12	NVIDIA Hifi-GAN 声码器（en-US）（Kong 等人，2020年） ； https://huggingface.co/nvidia/tts_hifigan
13	CommonVoice 14.0 英语上的 CRDNN https://huggingface.co/speechbrain/asr-crdnn-commonvoice-14-en
14	EnCodec（Défossez 等人，2023年）； https://huggingface.co/facebook/encodec_24千赫
15	Bark; https://huggingface.co/suno/bark; https://huggingface.co/erogol/bark/tree/main
#	**数据集**
16	ESD（Zhou 等人，2021年） ⋆ ⋆ ⋆⋆ https://hltsingapore.github.io/ESD/download.html

17	LibriSpeech（Panayotov 等人，2015年）：train-clean-100、train-clean-360、train-other-500 ⋆ ⋆ ⋆ ⋆ ⋆https://www.openslr.org/12
18	CREMA-D（Cao 等人，2014） ⋆⋆ https://github.com/CheyneyComputerScience/CREMA-D
19	RAVDESS（Livingstone 和 Russo，2018年） https://datasets.activeloop.ai/docs/ml/datasets/ravdess-dataset/ https://zenodo.org/records/1188976
20	VCTK (Veaux 等人，2019年) ⋆ ⋆ ⋆ ⋆ ⋆⋆ https://datashare.ed.ac.uk/handle/10283/2651; https://huggingface.co/datasets/vctk
21	SAVEE(Haq 等人，2009); http://kahlan.eps.surrey.ac.uk/savee/ https://www.kaggle.com/datasets/ejlok1/surrey-audiovisual-expressed-emotion-savee
22	EMO-DB (Burkhardt 等人，2005); http://emodb.bilderbar.info/download/
23	LJSpeech（Ito 和 Johnson，2017年）；https://keithito.com/LJ-Speech-Dataset/
24	Libri-light（Kahn 等人，2020年）（仅训练部分） ⋆⋆ https://github.com/facebookresearch/libri-light/blob/main/数据_preparation/README.md
25	VoxCeleb-1、2（Chung 等人，2018年） ⋆ ⋆ ⋆ https://www.robots.ox.ac.uk/~vgg/data/voxceleb/index.html#about
26	LibriTTS（Zen 等人，2019年）：train-clean-100、train-clean-360、train-other-500 ⋆ ⋆ ⋆⋆ https://openslr.org/60/
27	CMU-MOSEI（Zadeh 等人，2018年）；http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
28	MUSAN（Snyder 等人，2015年）；https://www.openslr.org/17/
29	RIR（Ko 等人，2017年）；https://www.openslr.org/28/
30	VGAF（Sharma 等人，2021年）（来自 EmotiW 挑战赛）；https://sites.google.com/view/emotiw2023 https://www.kaggle.com/datasets/amirabdrahimov/vgaf-dataset
31	MSP-Podcast（Lotfian 和 Busso，2019年） https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html
#	**带预训练模型的软件**
32	Resemblyzer：https://github.com/resemble-ai/Resemblyzer ⋆ 模型：https://github.com/resemble-ai/Resemblyzer/blob/master/resemblyzer/pretrained.pt
33	VITS（Kim 等人，2021年）；https://github.com/jaywalnut310/vits/ 模型：https://drive.google.com/drive/folders/1ksarh-cJf3F5eKJjLVWY0X1j1qsQqiS2
34	基于 VITS 预训练的 PIPER：https://github.com/rhasspy/piper/?tab=readme-ov-file 模型：https://huggingface.co/datasets/rhasspy/piper-checkpoints/tree/main
35	RVC-Project：https://github.com/RVC-Project 模型：https://huggingface.co/lj1995/VoiceConversionWebUI/tree/main
36	DISSC（Maimon 和 Adi，2023年b）；https://github.com/gallilmaimon/DISSC

B. 结果

表9 展示了各系统在开发集和测试数据上的性能指标。

图13 展示了归一化性能矩阵，其中所有系统均按总体得分排序（非官方排名）。该热力图呈现了多个语音匿名化系统在三项关键指标上的归一化性能对比（相对于原始数据结果的变化）： Δ非加权平均召回率， Δ词错误率，以及 Δ等错误率。数值归一化至 0 到 1 的范围，其中 1 表示最佳性能，0 表示最差性能。

Δ非加权平均召回率和 Δ词错误率指标在归一化过程中会取反，以体现数值越低表示效用越好（效用损失越少）；而 Δ等错误率则保持不变，以体现隐私越高。系统按照综合得分进行排名，该得分为三个指标归一化得分的平均值，热力图也据此排序。从红色到绿色的颜色编码直观地展示系统的优势与不足：绿色表示性能出色，红色表示性能较差。

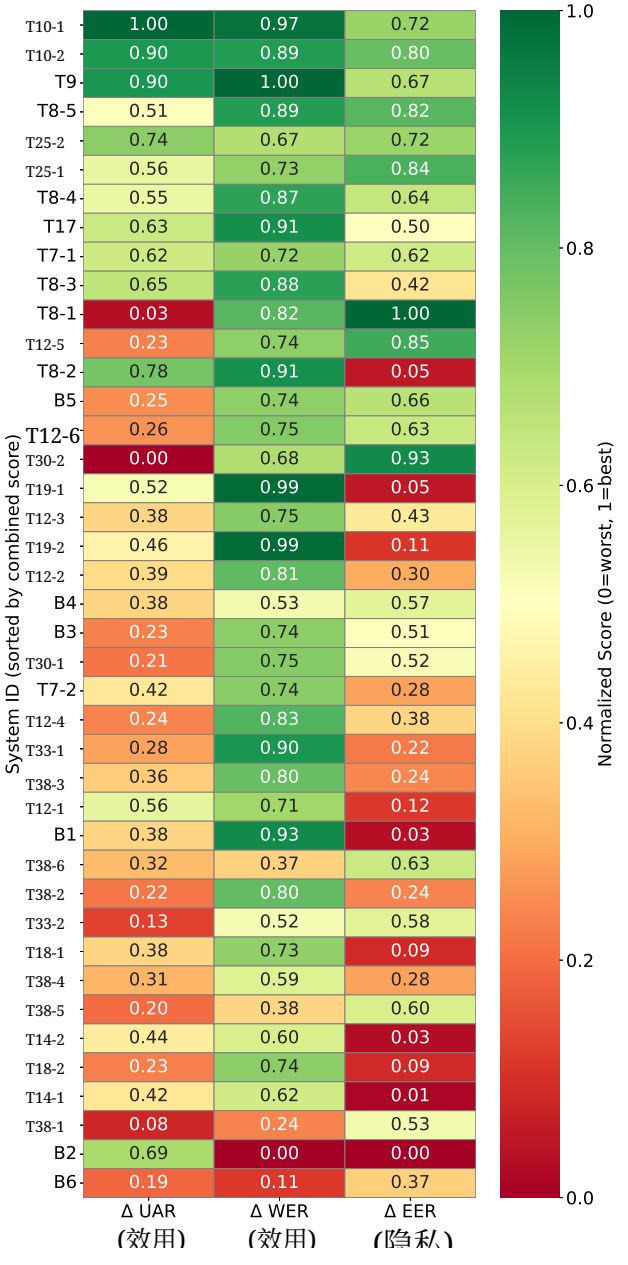


图13： 按总体得分排列的所有系统的归一化性能矩阵（非官方排名）。

表 9各系统的性能指标：*LibriSpeech* 上的等错误率（EER，%）和词错误率（WER，%），以及*IEMOCAP* 开发集和测试集上的非加权平均召回率（UAR，%）。

系统标识	开发集					test				
	等错误率，%			非加权平均召回率，%	词错误率，%	等错误率，%			非加权平均召回率，%	词错误率，%
	女性	male	aver			女性	male	aver		
orig	10.51	0.93	5.72	69.08	1.80	8.76	0.42	4.59	71.10	1.85
B1	10.94	7.45	9.20	42.71	3.07	7.47	4.68	6.08	42.80	2.91
B2	12.91	2.05	7.48	55.61	10.44	7.48	1.56	4.52	53.50	9.95
B3	28.43	22.04	25.24	38.09	4.29	27.92	26.72	27.32	37.60	4.35
B4	34.37	31.06	32.72	41.97	6.15	29.37	31.16	30.27	42.80	5.90
B5	35.82	32.92	34.37	38.08	4.73	33.95	34.73	34.34	38.20	4.37
B6	25.14	20.96	23.05	36.39	9.69	21.15	21.14	21.15	36.10	9.09
T7-1	35.23	27.64	31.43	48.70	4.86	34.67	30.23	32.45	51.22	4.47
T7-2	21.57	14.59	18.08	42.38	4.59	19.89	14.74	17.32	44.18	4.32
T8-1	47.33	48.12	47.72	30.08	3.74	48.90	50.15	49.53	30.59	3.75
T8-2	11.65	5.28	8.46	56.73	3.27	9.85	4.03	6.94	56.67	3.05
T8-3	25.00	16.90	20.95	51.28	3.30	21.19	25.39	23.29	52.13	3.29
T8-4	34.93	32.80	33.86	49.34	3.51	33.40	33.67	33.53	48.73	3.34
T8-5	39.63	40.84	40.24	47.08	3.45	42.50	40.05	41.28	47.10	3.20
T9	32.10	33.88	32.99	60.66	2.33	32.30	36.74	34.52	60.97	2.37
T10-1	43.33	34.32	38.82	65.98	2.57	37.24	36.56	36.90	64.45	2.57
T10-2	43.63	40.04	41.83	62.93	3.68	41.97	38.75	40.36	60.80	3.22
T12-1	19.18	6.99	13.08	49.20	4.97	12.04	7.39	9.71	49.12	4.60
T12-2	22.44	11.53	16.99	42.76	3.81	20.44	15.59	18.02	43.21	3.83
T12-3	26.14	18.32	22.23	44.67	4.21	25.53	22.54	24.03	42.78	4.29
T12-4	24.57	15.68	20.13	39.18	3.61	22.99	20.27	21.63	37.67	3.69
T12-5	43.32	44.10	43.71	38.06	4.76	43.61	41.88	42.75	37.60	4.36
T12-6	33.10	33.38	33.24	39.41	4.73	32.27	33.89	33.08	38.51	4.25
T14-1	19.32	3.42	11.37	45.45	6.14	7.12	3.16	5.14	44.23	5.28
T14-2	17.33	2.52	9.92	45.57	6.59	8.96	2.67	5.81	44.85	5.40
T17	36.08	41.44	38.76	51.71	2.95	34.13	20.26	27.20	51.41	3.02
T18-1	14.77	5.28	10.03	45.09	4.46	10.95	6.01	8.48	42.74	4.39
T18-2	13.78	4.81	9.29	43.52	4.79	10.77	6.04	8.40	37.37	4.34
T19-1	11.79	4.33	8.06	46.94	2.60	8.39	5.57	6.98	47.66	2.45
T19-2	16.62	7.01	11.82	43.20	2.48	11.11	7.76	9.43	45.50	2.44
T25-1	42.65	40.06	41.35	52.38	4.73	42.34	41.91	42.13	48.99	4.44
T25-2	36.65	35.72	36.18	55.31	5.19	37.74	35.85	36.79	55.39	4.85
T30-2	45.59	48.45	47.02	28.45	5.02	45.65	47.00	46.33	29.44	4.78
T30-1	29.10	22.20	25.65	38.48	4.13	30.11	25.39	27.75	36.78	4.28
T33-2	29.71	36.80	33.25	37.31	6.17	23.89	37.01	30.45	34.00	5.98
T33-1	22.16	13.35	17.76	38.69	3.16	14.23	14.69	14.46	39.27	3.12
T38-1	40.31	34.77	37.54	32.30	8.47	27.73	29.18	28.45	32.17	8.16
T38-2	25.43	15.84	20.63	37.86	4.18	13.51	16.97	15.24	36.98	3.86
T38-3	25.00	16.61	20.81	44.51	4.20	15.15	15.60	15.37	42.08	3.91
T38-4	27.70	21.90	24.80	42.51	5.85	16.03	18.48	17.26	40.45	5.51
T38-5	41.19	37.44	39.31	36.08	7.78	31.02	31.85	31.44	36.37	7.07
T38-6	42.58	38.35	40.47	41.19	7.80	31.91	33.63	32.77	40.79	7.11
T25-p1	45.12	41.22	43.17	53.84	4.72	44.19	41.40	42.80	50.07	4.45
T25-p2	38.29	34.11	36.20	57.86	4.98	37.94	35.51	36.73	55.79	5.12