无参考的语音性别模糊对抗方法

Yangyang Qu*, Michele Panariello*, Massimiliano Todisco* and Nicholas Evans*
* EURECOM, France

摘要—语音中的性别转换涉及数据收集带来的隐私风险，并且即使没有目标说话人的参考信息，输出中通常也会留下残留的性别特定线索。我们介绍了无参考对抗性别模糊化 RASO。创新包括一种基于性别的条件对抗学习框架，用于分离语言内容和与性别相关的声学标记，以及显式的正则化以使基频分布和共振峰轨迹与从性别平衡训练数据中学习到的中性特征保持一致。即使在半知情攻击模型下评估时，RASO 也显著保留了语言内容，并且比其他性别模糊化方法表现得更好。

I. 介绍

语音转换（VC）在隐私敏感应用中扮演着关键角色，例如在医疗场景中收集的语音数据匿名化 [1]。隐私保护涉及模糊特定说话人的特征（如声音、性别、年龄和口音），但保留其效用（如语言内容、自然度、韵律、情感和健康相关的线索）。本文所介绍的工作关注于模糊说话人的性别。[1] 传统语音转换方法依赖并行语料库或目标说话人参考 [4, 5]，面临两个根本性限制，即获取敏感目标语音数据的成本高昂以及在零样本场景下未能完全抑制区分性声学特征（如基频分布、共振峰轨迹）的残留线索可被重新识别攻击所利用 [6]。

为了应对这些挑战，我们提出了 RASO，一个基于 GAN 的无参考、性别中立的声音转换框架。我们的方法引入了以下关键创新：

1. 无参考、性别中立的转换通过条件对抗学习实现。我们的学习框架将与说话人无关的语言内容从区分性别的声学特征（基频 (F0) 分布和共振峰轨迹）中分离出来。一个判别器强制生成语音中的性别模糊，从而在不需要参考目标说话人的数据的情况下掩盖性别特定属性。

2. 显式声学正则化以实现分布中立性。为了确保性别中立，我们引入了一个性别特征修改模块，该模块对基频分布的概率密度和共振峰轨迹的时域动态范围进行归一化，使其与混合性别的语音统计相一致。此机制消除了声学参数中的性别特异性偏移，以实现人口层面的性别中立声学表示。

通过整合这些机制，RASO 提供了强大的解决方案，消除了对敏感目标说话人数据的需求，在有效抑制性别相关属性的同时保持了高语音可懂度和自然性，并通过将声学特征与混合性别的统计分布相匹配来确保群体级别的隐私。实验结果表明，RASO 超越了竞争的最先进方法 [6, 7]。

[1] 如 [2] 所示，性别指的是生物属性，而性别角色和行为是指社会构建的 [3]。

II. 相关工作

深度学习推动了语音转换的进步，基于 GAN 的方法如 CycleGAN-VC [4] 和 StarGANv2-VC [8] 通过循环一致性或风格编码在非平行、多域转换中解耦语言内容与说话人属性而引领该领域。这些模型擅长生成高保真的韵律细节，如音调轮廓、节奏和音色细微差别，但无意间保留了其表示中的隐私敏感的说话人线索（例如性别特定的形式素模式、声道特征），因为它们的设计优先考虑身份保存而非属性模糊。

在说话人匿名化的领域，近期的研究进展旨在平衡隐私保护与语言实用性。方等人（Fang et al.）[9] 引入了一种基础方法，通过融合说话人的 X-向量和神经波形模型，实现了身份模糊化同时保留了语言内容。在此基础上，斯里瓦斯塔瓦等人（Srivastava et al.）[10] 通过引入伪说话人选择策略对方法进行了改进，这些策略动态混合 X-向量以提升隐私和实用性的权衡。之后，冠军（Champion）[11] 提出了基于量化的方法来抑制声学特征中的说话人相关信息，其效果优于传统的噪声基方法。与此同时，帕纳里埃洛等人（Panariello et al.）[7] 采用了一种神经音频编解码策略，利用预训练的 EnCodec 和 Transformer 架构分离语义-声学标记用于合成。迈耶等人（Meyer et al.）[12] 进一步推进了该领域，通过使用生成对抗网络（GANs）生成伪嵌入来替换说话人身份，同时保持韵律细微差别。托马申科等人（Tomashenko et

al.) [1] 在半知情攻击者模型下评估了这些系统，强调需要标准化框架来评估多条件下的隐私-实用性权衡。

性别的模糊化研究较少。Stoidis 和 Cavallaro [13] 引入了 GenGAN，通过平滑频谱差异生成性别模糊的语音，实现了隐私和语音可懂度之间的平衡。Noé 等人 [6] 提出了一个使用对抗训练和归一化流的“零证据”框架，在分析/合成管道中抑制性别信息。Chouchane 等人 [14] 介绍了一个差异隐私的对抗自编码器框架，旨在通过缓解性别特定线索来保护语音生物特征中的性别信息。在他们另一项研究 [15] 中，他们分析了性别如何影响语音生物识别系统，并提出了减少与性别相关偏见的策略。Koutsogiannaki 等人 [16] 提出了一种方法，通过融合低频谱特征和韵律模式生成性别模糊的语音输出，以降低语音信号中性别的可辨别性。

III. 模型架构

为了在语音中实现性别模糊化，我们提出的框架采用了一种以隐私为导向的对抗架构，在抑制性别判别性声学特征的同时保留语言内容。该模型由两个核心组件组成：用于特征级别去标识化的生成器和多任务鉴别器。体系结构如图 1所示，并在下文进行描述。

A. 生成器：性别特征抑制网络

生成器旨在从输入语音中去除性别特定的声学标记，同时保持其他信息不变。它采用双分支架构，明确将语言内容与性别特征抑制分离。

1) 语言内容保存: 采用 Mel 谱图编码器来提取语言信息。输入的 Mel 谱图 $\boldsymbol{X} \in \mathbb{R}^{B\times 1\times 80\times T}$ 通过由下采样的残差块组成的分层编码模块压缩成一个潜在内容向量。输入的 Mel 谱图 $\boldsymbol{X} \in \mathbb{R}^{B\times 1\times 80\times T}$ 通过由下采样的残差块组成的分层编码模块压缩成一个潜在内容向量 $\boldsymbol{Z}_{\text{cont}} \in \mathbb{R}^{B\times C\times H\times W}$，其中 C 表示通道维度而 H, W 表示空间维度。

2) 性别特征修改: 三个专业模块被用来中和性别歧视的声学特征并保留语义内容：

共振峰操纵分支 - 该模块处理低频的 40 个梅尔带（$\boldsymbol{F}_{\text{low}} \in \mathbb{R}^{B\times 1\times 40\times T}$），以抑制性别区分的共振峰模式。通过引入性别条件嵌入机制，根据以下方式进行每个性别的共振峰编辑：

$$\boldsymbol{X}_{\text{mod}} = \boldsymbol{X} \odot \big(1 + \boldsymbol{W} \cdot \boldsymbol{s}(\boldsymbol{y}_{\text{org}})\big)$$

其中 $\boldsymbol{W} \in \mathbb{R}^{40\times 64}$ 是一个可学习的投影矩阵，$\boldsymbol{s}(\boldsymbol{y}_{\text{org}}) \in \mathbb{R}^{64}$ 是由输入性别标签 $\boldsymbol{y}_{\text{org}} \in \{0, 1\}$ 生成的嵌入向量（0 表示男性，1 表示女性）。标签 $\boldsymbol{s}(\boldsymbol{y}_{\text{org}})$ 通过条件嵌入层独立参数化，使模块能够应用特定于性别的频率调制策略。

对于女性输入（$\boldsymbol{y}_{\text{org}} = 1$），$\boldsymbol{s}(0)$ 被优化以增强高频段的衰减，中和特定于女性的形式音集中度。相反，对于男性输入（$\boldsymbol{y}_{\text{org}} = 0$），$\boldsymbol{s}(1)$ 针对低频带进行目标设定，以抑制男性主导的频谱特征。在训练过程中，$\boldsymbol{s}(\boldsymbol{y}_{\text{org}})$ 和 $\boldsymbol{W}$ 被联合优化。这种设计消除了对目标说话人参考的需求，仅依赖二元性别标签来实现方向性抑制，这有效地模糊了与性别相关的声学线索，同时保留了语言内容。

F0 中和分支 - 基本频率轮廓 f_0^{pred} 由模型 JDC [17] 预测。[2] 预测的基本频率轮廓 f_0^{pred} 通过对数域平移映射到性别中立的对应物 f_0^{shifted}：

$$f_0^{\text{shifted}} = \exp\left(\log(f_0^{\text{pred}}) + \log\left(\frac{\mu_{\text{neutral}}}{\bar{f}_0^{\text{org}}}\right)\right) \tag{1}$$

其中 $\bar{f}_0^{\text{org}}$ 表示输入语音的全局平均 F0，而 μ_{neutral} 在训练过程中通过指数移动平均进行更新：

$$\mu_{\text{neutral}}^{(t)} = \gamma\mu_{\text{neutral}}^{(t-1)} + (1-\gamma)\cdot \bar{f}_0^{\text{batch}} \tag{2}$$

其中 $\gamma = 0.99$ 和 $\bar{f}_0^{\text{batch}}$ 表示当前训练批次中所有语音样本的平均 F0，其作用是确保 μ_{neutral} 动态逼近混合性别训练语料库的全局 F0 统计数据同时抑制特定批次的变化。

特征融合与重构模块 - 性别特征抑制分支的输出，包括抑制定音低频 Mel 带和 F0 中性化轮廓，通过一个定音引导注意力机制与内容表示 $\boldsymbol{Z}_{\text{cont}}$ 融合，该机制从较低的 40 个 Mel 带提取与性别相关的光谱模式，并通过风格嵌入生成注意力图以突出性别中性频率区域。融合特征经过自适应实例归一化（AdaIN）[18] 的上采样残差块处理以恢复频谱分辨率，随后通过投影层重建 Mel 频谱图。这种设计抑制了性别区分的声学线索（定音偏移、F0 趋势），同时通过多尺度特征细化保留语言内容，从而实现高保真语音合成中的无参照性别模糊。

B. 判别器：对抗隐私变换

判别器 D 采用双目标架构，在对抗训练框架中执行两个互补的目标：保持语音可懂度和有效性别中立的语音生成。

[2] https://github.com/keums/melodyExtraction_JDC

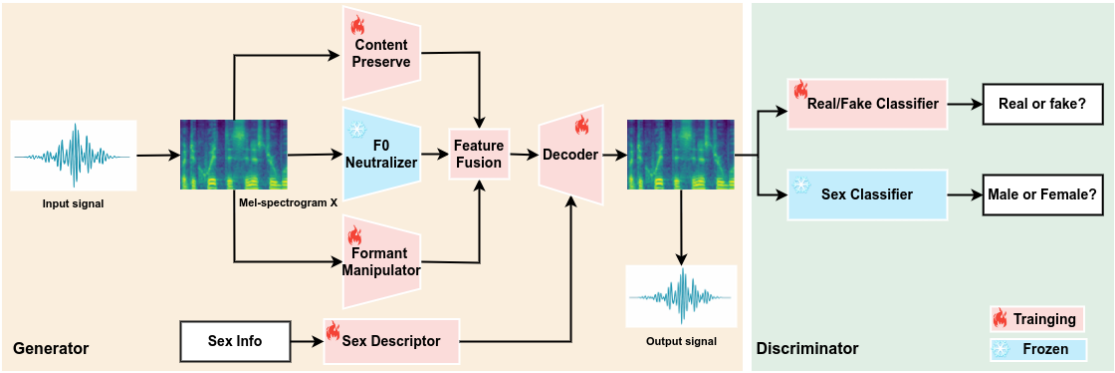

图 1. RASO 框架的架构。左侧显示生成器的训练过程。右侧显示判别器的训练过程。

1) 真实/伪造区分: 一个多尺度卷积网络与谱归一化被用于区分真实频谱图 $\boldsymbol{X}_{\text{real}}$ 和生成的频谱图 $\hat{\boldsymbol{X}}$。使用最小二乘生成对抗损失 [19] 来稳定训练。

2) 性别困惑歧视: 一个预训练的性别分类器，其参数被冻结，用于评估生成语音的性别模糊度 [20]。[3] 在训练过程中，判别器为生成器提供梯度以最大化分类器输出熵超过 $\hat{\boldsymbol{X}}$，而分类器参数保持不变以提供无偏评估。

C. 损失函数

我们的隐私驱动的损失框架通过多目标优化策略平衡性别模糊和语音可懂度。这是通过一组损失函数实现的，每个函数如下所述。

1) 对抗损失: 我们采用带有软标签的最小二乘生成对抗网络 (LSGAN) 损失 [19] 来稳定训练并促进频谱图结果 [21]:

$$\begin{aligned}\mathcal{L}_D^{\text{adv}} =& \frac{1}{2}\mathbb{E}_{\boldsymbol{X}_{\text{real}}\sim p_{\text{data}}}\left[(D(\boldsymbol{X}_{\text{real}})-0.95)^2\right] \\ &+\frac{1}{2}\mathbb{E}_{\hat{\boldsymbol{X}}\sim p_{\text{gen}}}\left[\left(D(\hat{\boldsymbol{X}})-0.05\right)^2\right], \end{aligned} \tag{3}$$

其中 $\boldsymbol{X}_{\text{real}}$ 表示真实的语音梅尔频谱图，$\hat{\boldsymbol{X}}$ 表示生成的性别中立语音，D 表示判别器。软标签（0.95 为真实，0.05 为假）相比硬标签（1 和 0）缓解了梯度消失问题。生成器的对抗损失由以下给出:

$$\mathcal{L}_G^{\text{adv}} = \frac{1}{2}\mathbb{E}_{\hat{\boldsymbol{X}}\sim p_{\text{gen}}}\left[\left(D(\hat{\boldsymbol{X}})-0.95\right)^2\right]. \tag{4}$$

2) 性别模糊损失: 为了强制性别中立性，我们最大化预训练性别分类器 C [20] 对生成语音的熵。损失定义为负香农熵:

$$\begin{aligned}\mathcal{L}_{\text{sex}} = -\mathbb{E}\Big[&\mathcal{P}_{\text{male}}(\hat{\boldsymbol{X}})\cdot\log\mathcal{P}_{\text{male}}(\hat{\boldsymbol{X}}) \\ &+\left(1-\mathcal{P}_{\text{male}}(\hat{\boldsymbol{X}})\right)\cdot\log\left(1-\mathcal{P}_{\text{male}}(\hat{\boldsymbol{X}})\right)\Big], \end{aligned} \tag{5}$$

其中 $\mathcal{P}_{\text{male}}(\hat{\boldsymbol{X}})\in[0,1]$ 是结果被分类为男性的概率。最小化 $\mathcal{L}_{\text{sex}}$ 强制执行 $\mathcal{P}_{\text{male}}\to 0.5$，确保均匀的类别分布。

3) 内容保留损失: 为了在转换过程中保留语言内容，我们使用一个预训练的自动语音识别（ASR）模型来实施特征级一致性损失 [22]。[4] 损失定义为:

$$\mathcal{L}_{\text{content}} = \mathbb{E}_{\boldsymbol{X}}\left[\left\|h_{\text{asr}}(\boldsymbol{X})-h_{\text{asr}}(\hat{\boldsymbol{X}})\right\|_1\right], \tag{6}$$

其中 $h_{\text{asr}}(\cdot)$ 表示来自 ASR 模型编码器的上下文特征提取器——一个捕捉语音中的音素和语义依赖关系的网络。这里，$\boldsymbol{X}$ 表示原始语音信号，$\hat{\boldsymbol{X}}$ 是变换后的输出，而 $\|\cdot\|_1$ 是 L1 范数，它最小化了原始和生成语音的高层特征之间的绝对差异。

4) 循环一致性损失: 为了在性别模糊过程中减轻内容退化，引入了循环一致性损失以确保原始语音和转换后语音之间的双向保真度。该损失定义为:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{\boldsymbol{X},s_{\text{src}}}\left[\left\|G\left(G(\boldsymbol{X},s_{\text{neutral}}),s_{\text{src}}\right)-\boldsymbol{X}\right\|_1\right], \tag{7}$$

其中 s_{src} 是由性别描述分支提取的源性别嵌入，而 s_{neutral} 是性别中性目标向量。通过最小化重建频谱图与原始频谱图之间的 L1 距离，该机制迫使生成器 G 学习一个可逆映射，在保持语言内容的同时中和性别特定的声学特征。

[3] https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender [4] https://github.com/yl4579/AuxiliaryASR

5) F0 中和损失: F0 被标准化为一个动态中性基线 μ_{neu}（初始值设为 150 Hz，即混合性别训练数据的中位数 F0），同时保持相对音高动态：

$$\mathcal{L}_{\text{F0}} = \mathbb{E}\Big[\big\| \bar{f}_0^{\text{gen}} - \mu_{\text{neu}} \big\|_1 + \lambda_{\text{rel}} \cdot \| \Delta\log(f_0^{\text{gen}}) - \Delta\log(f_0^{\text{org}}) \|_1 \Big], \quad (8)$$

其中 $\bar{f}_0^{\text{gen}}$ 和 $\bar{f}_0^{\text{org}}$ 分别表示生成语音和原始语音的平均 F0 值；$\Delta\log(f_0) = \log(f_0) - \log(\bar{f}_0)$ 表示捕捉相对动态的日志归一化音高轮廓；$\lambda_{\text{rel}} = 0.8$ 平衡绝对 F0 对齐与相对音高的保持。

6) 共鸣抑制损失: 生成的元音共振峰与混合性别的统计矩（均值 μ 和标准差 σ）对齐：

$$\mathcal{L}_{\text{formant}} = \sum_{k=1}^{3} \Big(\big\| \mu(\boldsymbol{F}_k^{\text{gen}}) - \mu(\boldsymbol{F}_k^{\text{neutral}}) \big\|_1 + \beta \cdot \big\| \sigma(\boldsymbol{F}_k^{\text{gen}}) - \sigma(\boldsymbol{F}_k^{\text{neutral}}) \big\|_1 \Big), \quad (9)$$

其中：$\boldsymbol{F}_k^{\text{gen}}$ 表示通过线性预测编码提取的生成语音的第 k 个元音共振峰；$\mu(\boldsymbol{F}_k^{\text{neutral}})$ 和 $\sigma(\boldsymbol{F}_k^{\text{neutral}})$ 分别是根据混合性别训练数据计算得到的所有元音共振峰的均值和标准差；$\beta = 0.3$ 控制元音共振峰的平滑度以平衡中立性和自然性。

7) 总生成器损失: 总生成器损失使用经验调整的权重来平衡性别模糊和语音可懂度：

$$\mathcal{L}_G = \alpha_1 \mathcal{L}_G^{\text{adv}} + \alpha_2 \mathcal{L}_{\text{sex}} + \alpha_3 \mathcal{L}_{\text{content}} \alpha_4 \mathcal{L}_{\text{F0}} + \alpha_5 \mathcal{L}_{\text{formant}} + \alpha_6 \mathcal{L}_{\text{cyc}}, \quad (10)$$

权重通过验证集进行网格搜索确定：$\alpha_1 = 1.0$,$\alpha_2 = 5.0$（优先考虑性别中立性），$\alpha_3 = 10.0$（内容保留的关键因素），$\alpha_4 = 2.0$,$\alpha_5 = 1.0$ 和 $\alpha_6 = 10.0$。

IV. 实验

A. 数据集

受之前关于语音隐私研究的启发 [1]，我们使用 LibriSpeech 数据集 [23] 进行实验，具体而言是用 train-clean-360 子集进行训练，并使用 test-clean 子集进行评估。训练集包含来自 921 名说话人的语音（482 名男性，439 名女性），而测试集包括 40 名未见过的说话人（20 名男性，20 名女性）。train-clean-360 子集中大规模、高质量的录音确保了模型训练的鲁棒性，而 test-clean 子集则提供了一个受控且未见过的数据集，用于严格评估隐私保护和转换质量。

表 I
不同攻击者场景下的性能比较
（EER↑ 表示更高的性别分类错误以实现更好的隐私；WER↓ 表示更低的语音识别错误以实现更好的可理解性）

模型类型	无知攻击者		半知情
	EER (%)↑	**WER (%) ↓**	**EER (%)↑**
Raw Data	7.22	1.84	–
Pan. et al. [7]	48.56	5.90	32.15
Noe et al. [6]	36.88	2.48	16.37
RASO	**55.38**	**2.47**	**47.25**

B. 训练详情

我们采用 AdamW 优化器 [24]，生成器的学习率为 10^{-5}，判别器的学习率为 10^{-4}。训练以 64 的批次大小，在 NVIDIA 3090 GPU 上使用 PyTorch 混合精度加速进行。基于验证损失应用了 150 个 epoch 的提前停止。

C. 目标度量标准

我们采用等错误率（EER）来评估性别分类，并使用词错误率（WER）来评估自动语音识别性能。EER 源自预训练的性别分类器 [20]，量化了性别特定声学特征的混淆，而 WER 依赖于在完整 LibriSpeech-train-960 数据集上训练的预训练 ASR 系统 [25] 评估语言内容的保留情况。

D. 评估

在语音隐私保护的背景下，我们对 RASO 的评估包含了两个最先进的基线，每个基线都根据其与性别模糊的关系进行了具体化。Noe 等人设计的 [6]，专门用于性别模糊，作为直接比较对象。补充这一点，Panariello 等人的 [7] 被纳入以基准测试相关方法。尽管他们的工作侧重于说话人匿名化，但也隐藏了与性别相关的特征。

为了模拟日益复杂的对抗场景，我们采用两种受到 VoicePrivacy Challenge [1] 启发的攻击模型。第一种是一个无知的攻击模型，假设攻击者不了解 RASO，并使用一个预训练的性别分类器 [5] 来对混淆后的语音进行分类。在第二种场景中，即半知情攻击 [26]，攻击者分别针对 Noe 等人 [6]、Panariello 等人 [7] 以及 RASO 生成的性别中性化数据集对性别分类器进行微调。这种设置评估了 RASO 对抗从竞争方法中适应混淆模式的分类器的韧性，提供了在不同框架之间的严格比较。

[5] https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender

我们的系统和两种竞争方法在表 I中分别展示了两种攻击模型的结果。还显示了原始（未经处理/未受保护）语音数据的结果。对于无知的攻击模型，RASO 达到了 55.38% 的 EER，显著优于两个竞争系统——Noe 等 [6] 的 36.88% 和 Panariello 等 [7] 的 48.56%。后者的结果表明，即使语音匿名化系统没有专门针对性别混淆进行调整，仍然可以有效工作，这很可能是因为转换中使用的原始/伪说话人的声音是随机性别的。RASO 维持了 2.47% 的 WER，与 Noe 等的 (2.48%) 相当，但远优于 Panariello 等的 (5.90%)。这些结果共同证明了成功抑制性别特定声学特征（例如共振峰模式、F0 曲线）和保持语言内容的有效性。

半知情攻击模型的结果显示出更加显著的差异，突显了我们方法的有效性。RASO 实现了 47.25%的等错误率（EER），远超竞争系统 32.15%和 16.37%的表现。这一重大改进强调了对抗训练和我们的多任务损失设计在应对更复杂攻击时赋予的鲁棒性，仍然无需访问目标说话人数据。在这两种攻击模型下，RASO 持续保持高 EER 和低 WER。

V. 结论

我们提出了一种综合对抗框架，用于在没有目标说话人参考的情况下实现鲁棒的性别模糊。我们的方法调整了共振峰模式和 F0 分布以中和语音中的性别线索，同时保持可理解性。实验结果证实了与竞争方法相比的进步，证明了我们在平衡性别信息模糊与语言内容保留方面的研究方法的价值。

在未来的研究中，一个潜在的扩展可能涉及引入控制性别模糊程度的机制，这将允许用户根据特定的隐私要求调整转换强度，从而增强框架在不同应用领域中的适应性。

参考文献

[1] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, "The voiceprivacy 2024 challenge evaluation plan," *arXiv preprint arXiv:2404.02677*, 2024.

[2] P.-G. Noé, M. Mohammadamini, D. Matrouf, T. Parcollet, A. Nautsch, and J.-F. Bonastre, "Adversarial disentanglement of speaker representation for attribute-driven privacy preservation," in *Interspeech 2021*, 2021, pp. 1902–1906.

[3] V. Prince, "Sex vs. gender," *International Journal of Transgenderism*, vol. 8, no. 4, pp. 29–32, 2005.

[4] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[5] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[6] P.-G. Noé, X. Miao, X. Wang, J. Yamagishi, J.-F. Bonastre, and D. Matrouf, "Hiding speaker' s sex in speech using zero-evidence speaker representation in an analysis/synthesis pipeline," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[7] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, "Speaker anonymization using neural audio codec language models," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4725–4729.

[8] Y. A. Li, A. Zare, and N. Mesgarani, "Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Proc. Interspeech 2021*, 2021, pp. 1349–1353.

[9] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[10] B. M. L. Srivastava, N. Tomashenko, X. Wang, E. Vincent, J. Yamagishi, M. Maouche, A. Bellet, and M. Tommasi, "Design choices for x-vector based speaker anonymization," *arXiv preprint*

arXiv:2005.08601, 2020.

[11] P. Champion, "Anonymizing speech: Evaluating and designing speaker anonymization techniques," *arXiv preprint arXiv:2308.04455*, 2023.

[12] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, "Prosody is not identity: A speaker anonymization approach using prosody cloning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] D. Stoidis and A. Cavallaro, "Generating gender-ambiguous voices for privacy-preserving speech recognition," *arXiv preprint arXiv:2207.01052*, 2022.

[14] O. Chouchane, M. Panariello, O. Zari, I. Kerenciler, I. Chihaoui, M. Todisco, and M. Önen, "Differentially private adversarial auto-encoder to protect gender in voice biometrics," in *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, 2023, pp. 127–132.

[15] O. Chouchane, M. Panariello, C. Galdi, M. Todisco, and N. Evans, "Fairness and privacy in voice biometrics: A study of gender influences using wav2vec 2.0," in *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2023, pp. 1–7.

[16] M. Koutsogiannaki, S. M. Dowall, and I. Agiomyrgiannakis, "Gender-ambiguous voice generation through feminine speaking style transfer in male voices," *arXiv preprint arXiv:2403.07661*, 2024.

[17] S. Kum and J. Nam, "Joint detection and classification of singing voice melody using convolutional recurrent neural networks," *Applied Sciences*, vol. 9, no. 7, p. 1324, 2019.

[18] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[19] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[20] F. Burkhardt, J. Wagner, H. Wierstorf, F. Eyben, and B. Schuller, "Speech-based age and gender prediction with transformers," in *Speech Communication; 15th ITG Conference*. VDE, 2023, pp. 46–50.

[21] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, 2016.

[22] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[24] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[25] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[26] B. M. L. Srivastava, N. Vauquier, M. Sahidullah, A. Bellet, M. Tommasi, and E. Vincent, "Evaluating voice conversion-based privacy protection against informed attackers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2802–2806.