Reference-free Adversarial Sex Obfuscation in Speech

Yangyang Qu*, Michele Panariello, Massimiliano Todisco and Nicholas Evans
EURECOM, France
E-mail:{quy, panariel, todisco, evans}@eurecom.fr

***Abstract*—Sex conversion in speech involves privacy risks from data collection and often leaves residual sex-specific cues in outputs, even when target speaker references are unavailable. We introduce RASO for Reference-free Adversarial Sex Obfuscation. Innovations include a sex-conditional adversarial learning framework to disentangle linguistic content from sex-related acoustic markers and explicit regularisation to align fundamental frequency distributions and formant trajectories with sex-neutral characteristics learned from sex-balanced training data. RASO preserves linguistic content and, even when assessed under a semi-informed attack model, it significantly outperforms a competing approach to sex obfuscation.**

I. Introduction

Voice Conversion (VC) plays a critical role in privacy-sensitive applications, e.g. anonymisation [1] of speech data collected in healthcare scenarios. Privacy preservation involves the obfuscation of speaker-specific traits (e.g., the voice, sex, age and accent) but the preservation of utility (e.g., the linguistic content, naturalness, prosody, emotion and health-related cues). The work presented in this paper concerns obfuscation of the speaker's sex.[1] Traditional voice conversion methods, which rely on parallel corpora or target speaker references [4], [5], face two fundamental limitations, namely the high cost of acquiring sensitive target speech data and the incomplete suppression of sex-discriminative acoustic features (e.g., fundamental frequency distributions, formant trajectories) in zero-shot scenarios, which leaves residual cues exploitable by re-identification attacks [6].

To address these challenges, we propose RASO, a GAN-based framework for reference-free, sex-neutral voice conversion. Our approach introduces the following key innovations:

1. Reference-free, sex-neutral conversion via conditional adversarial learning. Our learning framework disentangles speaker-agnostic linguistic content from sex-discriminative acoustic features (fundamental frequency(F0) distributions and formant trajectories). A discriminator enforces sex ambiguity in generated speech, enabling the obfuscation of sex-specific attributes without requiring reference target speaker data.

2. Explicit acoustic regularisation for distributional neutrality. To ensure sex neutrality, we introduce a sex feature modification module that normalises the probability density of fundamental frequency distributions and the temporal dynamic range of formant trajectories to align with mixed-sex speech statistics. This mechanism eliminates sex-specific offsets in acoustic parameters to achieve population-level, sex-neutral acoustic representations.

By integrating these mechanisms, RASO offers a robust solution which eliminates the need for sensitive target speaker data, effectively suppresses sex-related attributes while maintaining high speech intelligibility and naturalness, and ensures population-level privacy by aligning acoustic features with mixed-sex statistical distributions. Experimental results show that RASO surpasses competing state-of-the-art methods [6], [7].

II. Related Work

Deep learning has propelled voice conversion advancements, with GAN-based methods like CycleGAN-VC [4] and StarGANv2-VC [8] leading the field by disentangling linguistic content from speaker attributes via cycle-consistency or style encoding for non-parallel, multi-domain conversion. These models excel at generating high-fidelity prosodic details, such as pitch contours, rhythm, and timbral nuances, but inadvertently retain privacy-sensitive speaker cues (e.g., sex-specific formant patterns, vocal tract characteristics) embedded in their representations, as their design prioritises identity preservation over attribute obfuscation.

In the realm of speaker anonymisation, recent advancements have sought to balance privacy preservation with linguistic utility. Early work by Fang et al. [9] introduced a foundational approach by fusing speaker X-vectors with neural waveform models, enabling identity obfuscation while retaining linguistic content. Building on this, Srivastava et al. [10] refined the approach by introducing pseudo-speaker selection strategies that dynamically mixed X-vectors to enhance privacy-utility trade-offs. Later, Champion [11] proposed quantisation-based transformations to suppress speaker-related information in acoustic features, outperforming traditional noise-based methods. Concurrently, Panariello et al. [7] adopted a neural audio codec strategy, leveraging pre-trained EnCodec and Transformer architectures to disentangle semantic-acoustic tokens for synthesis. Meyer et al. [12] further advanced the field by generating pseudo-embeddings via GANs to replace speaker identities while preserving prosodic nuances. Tomashenko et al. [1] evaluated these systems under semi-informed attacker

* Corresponding author.

This work was supported by the French Agence Nationale de la Recherche (ANR) via the SpeechPrivacy (ANR-23-CE23-0022) project.

[1]As in [2], sex refers to biological attributes, whereas gender refers to socially constructed roles and behaviour [3].

models, emphasising the need for standardised frameworks to assess multi-condition privacy-utility trade-offs.

There is less work in sex obfuscation. Stoidis and Cavallaro [13] introduced GenGAN, which generates sex-ambiguous speech by smoothing spectral differences, achieving balanced privacy and speech intelligibility. Noé et al. [6] propose a "zero-evidence" framework using adversarial training and normalising flows to suppress sex information in an analysis/synthesis pipeline. Chouchane et al. [14] present a differentially private adversarial auto-encoder framework, designed to protect sex information in voice biometrics by mitigating sex-specific cues. In their other work [15], they analysed how sex affects voice biometric systems and proposed strategies to reduce sex-related biases. Koutsogiannaki et al. [16] propose a method that blends low-frequency spectral characteristics with prosodic patterns to generate sex-ambiguous speech outputs and reduce the discriminability of sex attributes in speech signals.

III. Model Architecture

To achieve sex obfuscation in speech, our proposed framework employs a privacy-driven adversarial architecture that suppresses sex-discriminative acoustic features while preserving linguistic content. The model consists of two core components: a generator for feature-level de-identification and a multi-task discriminator. The architecture is illustrated in Fig. 1 and is described in the following.

A. Generator: sex feature suppression network

The generator aims to remove sex-specific acoustic markers from input speech while keeping other information intact. It employs a dual-branch architecture which explicitly decouples linguistic content from sex feature suppression.

1) Linguistic content preservation: A Mel-spectrogram encoder is employed to extract linguistic information. The input Mel-spectrogram $\mathbf{X} \in \mathbb{R}^{B\times 1\times 80\times T}$ is compressed into a latent content vector by a hierarchical encoding module consisting of residual blocks with downsampling. The input Mel-spectrogram $\mathbf{X} \in \mathbb{R}^{B\times 1\times 80\times T}$ is compressed into a latent content vector $\mathbf{Z}_{\text{cont}} \in \mathbb{R}^{B\times C\times H\times W}$ by a hierarchical encoding module consisting of residual blocks with downsampling, where C denotes the channel dimension and H, W represent spatial dimensions.

2) Sex feature modification: Three specialised modules are employed to neutralise sex-discriminative acoustic features and preserve semantic content:

Formant Manipulation Branch - This module processes the lower 40 Mel bands ($\mathbf{F}_{\text{low}} \in \mathbb{R}^{B\times 1\times 40\times T}$) to suppress sex-discriminative formant patterns. By introducing a sex-conditioned embedding mechanism, the formants for each sex are edited according to:

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

where $\mathbf{W} \in \mathbb{R}^{40\times 64}$ is a learnable projection matrix, and $\mathbf{s}(\mathbf{y}_{\text{org}}) \in \mathbb{R}^{64}$ is an embedding vector generated by the sex descriptor based on the input sex label $\mathbf{y}_{\text{org}} \in \{0, 1\}$ (0 represents male, 1 represents female). Label $\mathbf{s}(\mathbf{y}_{\text{org}})$ is parameterised independently via a conditional embedding layer, enabling the module to apply sex-specific frequency modulation strategies.

For female inputs ($\mathbf{y}_{\text{org}} = 1$), $\mathbf{s}(0)$ is optimised to enhance attenuation at higher frequencies to neutralise female-specific formant concentrations. Conversely, for male inputs ($\mathbf{y}_{\text{org}} = 0$), $\mathbf{s}(1)$ targets low-frequency bands to suppress male-dominant spectral features. During training, $\mathbf{s}(\mathbf{y}_{\text{org}})$ and $\mathbf{W}$ are jointly optimised. This design eliminates the need for target speaker references, relying solely on binary sex labels to achieve directional suppression, which effectively obfuscates sex-related acoustic cues while preserving linguistic content.

F0 Neutralization Branch - The fundamental frequency contour f_0^{pred} is predicted by the model JDC [17]. [2] The predicted fundamental frequency contour f_0^{pred} is mapped to the sex-neutralized counterpart f_0^{shifted} via log-domain shifting:

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

where $\bar{f}_0^{\text{org}}$denotes the global mean F0 of the input speech and where μ_{neutral} is updated via exponential moving average during training:

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

where $\gamma = 0.99$ and $\bar{f}_0^{\text{batch}}$ represent the average F0 across all speech samples in the current training batch, which acts to ensure that μ_{neutral} dynamically approximates the global F0 statistics of the mixed-sex training corpus while suppressing batch-specific fluctuations.

Feature Fusion and Reconstruction Module - The outputs of the sex feature suppression branch, including formant-suppressed low-frequency Mel-bands and F0-neutralised contours, are fused with the content representation $\mathbf{Z}_{\text{cont}}$ via a formant-guided attention mechanism which extracts sex-relevant spectral patterns from the lower 40 Mel-bands and generates attention maps through style embeddings to highlight sex-neutral frequency regions. Fused features are processed through upsampling residual blocks with adaptive instance normalisation (AdaIN)[18] to restore spectral resolution, followed by a projection layer to reconstruct the Mel-spectrogram. This design suppresses sex-discriminative acoustic cues (formant shifts, F0 trends) while preserving linguistic content through multi-scale feature refinement, enabling reference-free sex obfuscation with high-fidelity speech synthesis.

B. Discriminator: Adversarial Privacy Transformation

The discriminator D is designed with a dual-objective architecture to enforce two complementary objectives in the adversarial training framework: speech generation with preserved speech intelligibility and effective sex neutrality.

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

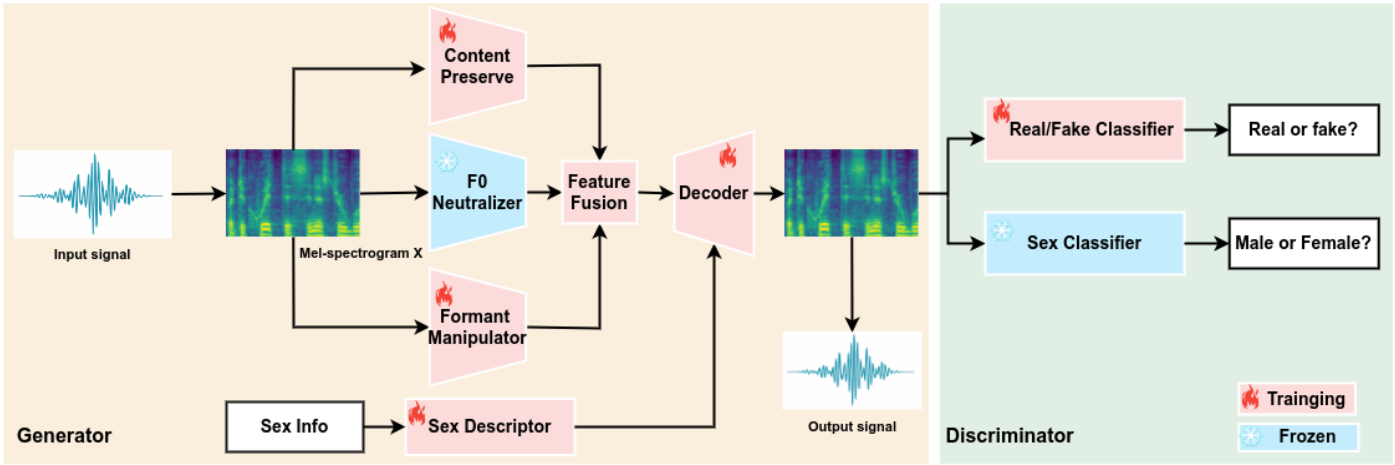

Fig. 1. Architecture of the RASO framework. The left side shows the training process of the generator. The right side shows the training process of the discriminator.

1) Real/Fake Discrimination: A multi-scale convolutional network with spectral normalisation is used to distinguish between real spectrograms $\mathbf{X}_{\text{real}}$ and generated spectrograms $\hat{\mathbf{X}}$. A least-squares generative adversarial loss [19] is used to stabilise training.

2) Sex Confusion Discrimination: A pre-trained sex classifier with frozen parameters is used to evaluate the sex ambiguity of generated speech [20].[3] During training, the discriminator provides gradients to the generator to maximise the classifier output entropy over $\hat{\mathbf{X}}$, while the classifier parameters remain fixed to provide an unbiased evaluation.

C. Loss Functions

Our privacy-driven loss framework balances sex obfuscation and speech intelligibility through a multi-objective optimisation strategy. This is achieved using a set of loss functions, each of which is described below.

1) Adversarial Loss: We adopt the Least Squares GAN (LSGAN) loss [19] with soft labels to stabilise training and promote spectrogram results[21]:

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

where $\mathbf{X}_{\text{real}}$ denotes real speech Mel-spectrograms, $\hat{\mathbf{X}}$ denotes generated sex-neutral speech, and D denotes the discriminator. The soft labels (0.95 for real, 0.05 for fake) mitigate gradient vanishing compared to hard labels (1 and 0). The adversarial loss of the generator is given by:

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

2) Sex Ambiguity Loss: To enforce sex neutrality, we maximise the entropy of a pre-trained sex classifier C [20] over generated speech. The loss is defined as the negative Shannon entropy:

$$\mathcal{L}_{\text{sex}} = -\mathbb{E}\Big[\mathcal{P}_{\text{male}}(\hat{\mathbf{X}}) \cdot \log \mathcal{P}_{\text{male}}(\hat{\mathbf{X}}) + \left(1 - \mathcal{P}_{\text{male}}(\hat{\mathbf{X}})\right) \cdot \log\left(1 - \mathcal{P}_{\text{male}}(\hat{\mathbf{X}})\right)\Big], \quad (5)$$

where $\mathcal{P}_{\text{male}}(\hat{\mathbf{X}}) \in [0, 1]$ is the probability that the outcome is classified as male. Minimizing $\mathcal{L}_{\text{sex}}$ forces $\mathcal{P}_{\text{male}} \to 0.5$, ensuring a uniform class distribution.

3) Content Preservation Loss: To ensure linguistic content is retained during transformation, we employ a feature-level consistency loss using a pre-trained automatic speech recognition (ASR) model [22].[4] The loss is defined as:

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

where $h_{\text{asr}}(\cdot)$ denotes the contextual feature extractor from an ASR model encoder—a network which captures phonetic and semantic dependencies in speech. Here, $\mathbf{X}$ represents the original speech signal, $\hat{\mathbf{X}}$ is the transformed output, and $\|\cdot\|_1$ is the L1 norm, which minimizes the absolute difference between high-level features of the original and generated speech.

4) Cycle Consistency Loss: To mitigate content degradation during sex obfuscation, a cycle consistency loss is introduced to enforce bidirectional fidelity between the original and transformed speech. The loss is defined as:

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

where s_{src} is the source sex embedding extracted by the sex descriptor branch, and s_{neutral} is the sex-neutral target vector. By minimising the L1 distance between reconstructed and original spectrograms, this mechanism forces the generator G to learn an invertible mapping, preserving linguistic content while neutralising sex-specific acoustics.

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

[4]https://github.com/yl4579/AuxiliaryASR

5) F0 Neutralization Loss: The F0 is normalised to a dynamic neutral baseline μ_{neu} (initialised at 150 Hz, the median F0 of mixed-sex training data) while preserving relative pitch dynamics:

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

where $\bar{f}_0^{\text{gen}}$ and $\bar{f}_0^{\text{org}}$ denote the mean F0 values of generated and original speech, respectively; $\Delta\log(f_0) = \log(f_0) - \log(\bar{f}_0)$ represents log-normalized pitch contours capturing relative dynamics; and $\lambda_{\text{rel}} = 0.8$ balances absolute F0 alignment and relative pitch preservation.

6) Formant Suppression Loss: Generated formants are aligned with mixed-sex statistical moments (mean μ and standard deviation σ):

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

where: $\mathbf{F}_k^{\text{gen}}$ denotes the k-th formant of generated speech (extracted via Linear Predictive Coding); $\mu(\mathbf{F}_k^{\text{neutral}})$ and $\sigma(\mathbf{F}_k^{\text{neutral}})$ are the global mean and standard deviation of formants computed from mixed-sex training data; $\beta = 0.3$ controls formant smoothness to balance neutrality and naturalness.

7) Total Generator Loss: The total generator loss is used with empirically tuned weights to balance sex obfuscation and speech intelligibility:

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

with weights determined from a grid search using a validation set: $\alpha_1 = 1.0$, $\alpha_2 = 5.0$ (prioritising sex neutrality), $\alpha_3 = 10.0$ (critical for content preservation), $\alpha_4 = 2.0$, $\alpha_5 = 1.0$, and $\alpha_6 = 10.0$.

IV. Experiments

A. Dataset

Inspired by previous research on voice privacy[1], we employ the LibriSpeech corpus [23] for experiments, specifically the train-clean-360 subset for training and test-clean subset for evaluation. The training set contains speech from 921 speakers (482 male, 439 female), while the test set includes 40 unseen speakers (20 male, 20 female). The large-scale, high-quality recordings in the train-clean-360 subset ensure robust model training, while the test-clean subset provides a controlled, unseen dataset for rigorous privacy and conversion quality assessment.

TABLE I
Performance Comparison Under Different Attacker Scenarios
(EER↑ indicates higher sex classification error for better privacy; WER↓ indicates lower speech recognition error for better intelligibility)

Model Type	Ignorant Attacker		Semi Informed
	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. Training Details

We employ the AdamW optimizer [24] with learning rates of 10^{-5} for the generator and 10^{-4} for the discriminator. Training is performed with a batch size of 64 and an NVIDIA 3090 GPU with PyTorch mixed-precision acceleration. Early stopping based on the validation loss is applied for 150 epochs.

C. Objective Metrics

We adopt the Equal Error Rate (EER) to evaluate sex classification and the Word Error Rate (WER) to evaluation ASR performance. The EER is derived from the pre-trained sex classifier [20] and quantifies the obfuscation of sex-specific acoustic features, while the WER relies on a pre-trained ASR system [25] trained on the full LibriSpeech-train-960 dataset to assess the preservation of linguistic content.

D. Evaluations

Within the context of voice privacy protection, our evaluation of RASO incorporates two state-of-the-art baselines, each contextualised by their relationship to sex obfuscation. Noe et al. [6], explicitly designed for sex obfuscation, serves as a direct comparator. Complementing this, Panariello et al. [7] is included to benchmark a related approach. Although their work focuses on speaker anonymisation, it hides the speakerID while also hiding sex-related features.

To simulate adversarial scenarios of increasing sophistication, we adopt two attack models inspired by the VoicePrivacy Challenge [1]. The first, an ignorant attack model, assumes the attacker lacks knowledge of RASO and uses a pre-trained sex classifier [5] to classify obfuscated speech. In the second scenario, a semi-informed attack [26], the attacker fine-tunes a sex classifier on sex-neutralised datasets generated by Noe et al. [6], Panariello et al. [7] and RASO, respectively. This setup assesses RASO's resilience against classifiers adapted to obfuscation patterns from competing methods, providing a rigorous comparison across frameworks.

Results for our system and two competing methods are presented in Table I for both attack models. Also shown are results for raw (unprocessed/unprotected) speech data. For the ignorant attack model, RASO achieves an EER of 55.38%, significantly outperforming results for both competing systems – 36.88% for Noe et al. [6] and 48.56% for Panariello et al. [7].

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

The latter result shows that even voice anonymisation systems, though not tuned specifically for sex obfuscation, can still be effective, most likely because target/pseudo speaker voices used in the conversion are of random sex. RASO maintains a WER of 2.47%, comparable to that of Noe et al. (2.48%) but far superior to that of Panariello et al. (5.90%). Together, these results demonstrate the successful suppression of sex-specific acoustic features (e.g., formant patterns, F0 contours) and the preservation of linguistic content.

The results for the semi-informed attack model exhibit even more pronounced disparities, underscoring the efficacy of our approach. RASO achieves an EER of 47.25%, significantly outperforming the 32.15% and 16.37% for competing systems. This substantial improvement highlights the resilience conferred by adversarial training and our multi-task loss design against more sophisticated attacks, still without access to target speaker data. Across both attack models, RASO consistently attains a high EER and low WER.

V. Conclusions

We propose an integrated adversarial framework for robust sex obfuscation without target speaker references. Our approach adjusts formant patterns and F0 distributions to neutralise sex cues in speech while preserving intelligibility. Experimental results confirm improvements over competing methods, demonstrating the merit of our approach in balancing the obfuscation of sex information with the preservation of linguistic content.

In future research, a potential extension could involve introducing mechanisms to control the degree of sex obfuscation, which would allow users to tailor the conversion intensity according to specific privacy requirements, thereby enhancing the framework's adaptability across diverse application domains.

References

[1] N. Tomashenko, X. Miao, P. Champion, *et al.*, "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, *et al.*, "Speaker anonymization using x-vector and neural waveform models," *arXiv preprint arXiv:1905.13561*, 2019.

[10] B. M. L. Srivastava, N. Tomashenko, X. Wang, *et al.*, "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, *et al.*, "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, *et al.*, *SpeechBrain: A general-purpose speech toolkit*, arXiv:2106.04624, 2021. arXiv: 2106 . 04624 [eess.AS].

[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.