Expert Systems With Applications 304 (2026) 130670

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

# SAM: A privacy-preserving framework for selective attribute masking in voice recordings

Anil Pudasaini [a,*], Muna Al-Hawawreh [a], Mohamed Reda Bouadjenek [a], Hakim Hacid [b], Sunil Aryal [a]

[a] School of Information Technology, Deakin University, Waurn Ponds Campus, 3216, Geelong, VIC, Australia
[b] Technology Innovation Institute, United Arab Emirates

ARTICLE INFO

*Keywords:*
Audio profiling
Speaker profiling
Privacy preservation
Age detection
Gender detection
Accent detection
Selective masking
Adversarial learning

ABSTRACT

Human voice is a rich source of information that can reveal a range of sensitive personal attributes, such as age, gender, and country of origin. With advances in Artificial Intelligence (AI), especially in speech processing, these personal attributes can now be inferred on a scale with high accuracy, raising serious privacy concerns. In fact, the ability to extract demographic or identification information from voice data poses risks related to surveillance, profiling, and misuse of personal data, highlighting the urgent need for privacy-preserving solutions in voice-based AI systems. Therefore, this paper proposes **Selective Attribute Masking (SAM)**, a new model-agnostic framework that uses gradient-based adversarial perturbations to suppress the inference of specific speaker attributes from voice recordings, while preserving the accuracy of non-target attributes and maintaining the utility of Automatic Speech Recognition (ASR). Experimental results using CommonVoice dataset demonstrate that SAM achieves selective masking success rates of up to **74.5 %** for age, **59.4 %** for gender, and **54.6 %** for accent–substantially outperforming baseline methods. At the same time, voice utility (that is, ASR) remains largely unaffected, with the word error rate increasing by less than **3 % absolute** under moderate perturbations. These findings demonstrate the effectiveness of our proposed framework (SAM) in balancing privacy and utility in voice-based systems.

## 1. Introduction

With the advancement of conversational Human-Computer interaction, voice has emerged as a primary modality for interacting with systems and intelligent agents. From virtual assistants to voice-enabled applications, speech-based interfaces are now embedded in everyday technologies, offering natural, intuitive, and hands-free communication between humans and machines. Typically, users provide voice input for specific tasks–such as issuing spoken commands–under the assumption that only the content of their speech will be processed. However, the human voice encodes much more than linguistic information; even brief recordings can inadvertently reveal sensitive personal attributes such as age, gender, emotional state, or ethnic background (Singh, 2019). For example, Amazon has patented technology for inferring users' physical and emotional characteristics from voice, such as detecting illness or stress (Aloufi et al., 2019). Integrated into assistants like *Alexa*®, such systems could enable personalized content delivery–like health-related ads during illness or calming products during stress–raising significant concerns about implicit profiling and persuasive targeting. These secondary inferences highlight the growing privacy risks in voice-driven applications.

Conventional privacy-preserving mechanisms for voice data, such as anonymization or voice conversion, can mitigate the risk of sensitive attribute inference. However, these approaches are often coarse-grained and degrade the utility of the speech signal, negatively impacting downstream tasks like automatic speech recognition or speaker verification (Cai et al., 2024; Tomashenko et al., 2022). For example, one could avoid sharing raw audio or heavily distorting it, which degrades the quality and usefulness of the spoken content (Chen et al., 2024). Similarly, techniques such as masking certain frequency ranges in the voice signal can obscure identifying characteristics, but this also compromises speech intelligibility (i.e., how clearly the speech can be understood by a listener or system) and hinders the effectiveness of voice-based applications.

However, a key limitation of these approaches is that they operate in an all-or-nothing manner, anonymizing the entire signal rather than

* Corresponding author.
*E-mail addresses:* s223786275@deakin.edu.au (A. Pudasaini), muna.alhawawreh@deakin.edu.au (M. Al-Hawawreh), reda.bouadjenek@deakin.edu.au (M.R. Bouadjenek), hakim.hacid@tii.ae (H. Hacid), sunil.aryal@deakin.edu.au (S. Aryal).

https://doi.org/10.1016/j.eswa.2025.130670
Received 26 September 2025; Received in revised form 15 November 2025; Accepted 30 November 2025
Available online 7 December 2025

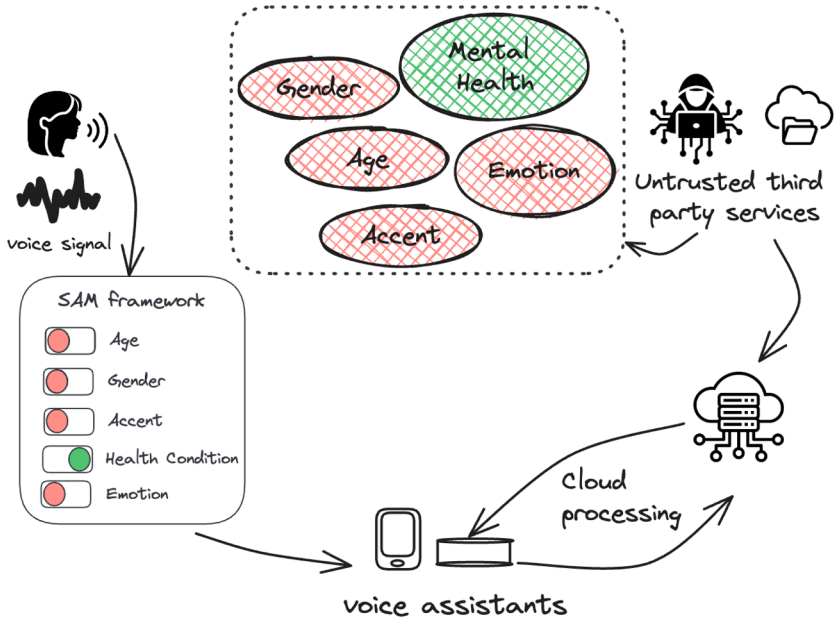



**Fig. 1.** Overview of the proposed privacy-preserving framework.

giving users control over which attributes to conceal. In practice, attributes such as gender, accent, or age may differ in sensitivity depending on the context, making selective, attribute-specific protection a more realistic requirement. As a result, existing methods often skew the privacy–utility balance too far toward one side—either protecting attributes at the expense of intelligibility or preserving intelligibility while leaving attributes exposed. What is needed are solutions that achieve a more balanced trade-off, enabling fine-grained control that masks sensitive attributes without unduly compromising system performance.

To address these limitations and enable fine-grained control over which voice attributes are shared, we propose a new framework called **Selective Attribute Masking (SAM)**. As shown in Fig. 1, SAM allows users to selectively protect specific voice-based attributes by choosing which features they wish to conceal before sharing their speech data. This framework offers a more fine-grained method of privacy protection without compromising the utility of voice interactions. It uses targeted adversarial perturbations that suppress the inference of chosen attributes while preserving the intended voice content, such as commands or transcription. This enables a flexible privacy-preserving mechanism that operates before cloud transmission. SAM assumes a white-box threat model (Goodfellow et al., 2015), where the user (defender) has complete knowledge of the attribute inference model. Importantly, SAM is model-agnostic and can be applied to any differentiable attribute classifier, making it adaptable to various deployment scenarios. By computing gradients through this network, SAM generates minimal adversarial perturbations on input speech spectrograms to mask a chosen target attribute. The perturbation is applied such that the target attribute's predicted class is flipped (or confidence reduced) while the other non-target attributes' predictions are ideally preserved. To the best of our knowledge, this is the first study to demonstrate selective masking of multiple voice attributes using adversarial attacks on real-world speech data. The key contributions of this work are summarized as follows:

1. We propose **Selective Attribute Masking (SAM)**, a new framework for fine-grained privacy control over speaker attributes in voice data using adversarial learning.
2. We propose a consistency-based composite loss function for selective attribute masking that maximizes misclassification of a target attribute while preserving non-target predictions. The formulation generalizes to multi-head models with an arbitrary number of attributes.
3. We introduce a new metric – *selective masking success rate* – to quantify privacy by measuring how often the target attribute is successfully hidden without degrading non-target attributes.
4. Finally, we evaluate the utility trade-off by measuring ASR performance on adversarially perturbed audio, showing minimal WER degradation under moderate perturbations.

## 2. Literature review

In this section, we first highlight the risks of inadvertent disclosure of sensitive voice attributes, then explore the use of adversarial machine learning for privacy protection, and finally discuss the limitations of current privacy methods.

### 2.1. Voice attributes and privacy threats

Speech signal carries a wealth of paralinguistic information that can be exploited beyond a user's intent (Pudasaini et al., 2025). Prior studies have documented that even "benign" voice inputs allow machine learning models to infer sensitive traits like the speaker's demographics, emotion (Kröger et al., 2020), or health conditions (Ali et al., 2019; Mohammed et al., 2023). This has spurred research into countermeasures that can protect users from unintended information leakage. A conventional line of defense is voice anonymization (Zhang et al., 2023), which aims to remove or hide personally identifiable characteristics (usually focusing on the identity of the speaker) before data are shared. The Voice Privacy Challenge series, formed in 2020, has advanced speaker anonymization techniques, often using voice conversion or speech synthesis methods to conceal identity while preserving linguistic content (Tomashenko et al., 2024). However, these methods do not offer selective control over which attributes are hidden. Furthermore, anonymization efforts have centered mostly on identity, leaving other attributes less explored.

### 2.2. Adversarial techniques for voice privacy

Adversarial learning has gained substantial attention in computer vision, where carefully crafted perturbations can mislead deep neural

classifiers without visibly altering input images. Building on the foundational work of Goodfellow et al. (2015), subsequent studies have examined both the creation and mitigation of such adversarial examples. In particular, Muthalagu et al. (2025) conducted a comprehensive evaluation of evasion attacks and defense mechanisms, demonstrating that even robust vision models remain vulnerable to adaptive perturbations. Although adversarial techniques originated and are most extensively studied in the image domain, researchers have extended them to speech for privacy-preserving applications. Unlike voice-conversion-based anonymization, adversarial methods add subtle perturbations to the input signal with the goal of misleading specific classifiers (Rabhi et al., 2024). The Fast Gradient Sign Method (FGSM) and its iterative extension, Projected Gradient Descent (PGD) (Madry et al., 2019), are commonly used to generate such gradient-based perturbations. Chen et al. (2024) applied iterative FGSM to perturb audio and prevent personalized speech generators (e.g., YourTTS Casanova et al., 2023) from cloning a speaker's voice. Similarly, psychoacoustic masking has been used to craft inaudible perturbations that fool x-vector-based speaker recognition systems (Wang et al., 2020), while Wang et al. (2024) proposed an "asynchronous" approach that imperceptibly perturbs speaker embeddings to disrupt verification systems without affecting human perception. Beyond speaker identity, Testa et al. (2023) introduced DARE-GP to mask emotional cues using genetic programming, and Jaiswal and Provost (Jaiswal & Provost, 2019) adversarially removed demographic traits from intermediate features. Collectively, these works demonstrate that adversarial noise can effectively conceal specific speaker attributes while preserving utility.

### 2.3. Gaps in multi-attribute privacy

Despite promising progress in voice privacy, much of the existing work focuses on masking a single sensitive attribute at a time. For example, several studies train dedicated models to suppress only speaker identity (Chen et al., 2022; Patino et al., 2021), emotion (Aloufi et al., 2019; Testa et al., 2023), or gender (Chouchane et al., 2023; Stoidis & Cavallaro, 2022). These attribute-specific approaches often rely on generative or domain-adversarial training, where the feature extractor is optimized to "forget" a particular attribute. While effective in narrow settings, these methods are not designed to scale to multiple concurrent privacy goals, and naively combining them can lead to interference or utility degradation.

There is a growing recognition of the need for multi-attribute privacy solutions that can target several attributes jointly or offer user-selectable privacy settings (Chen et al., 2024). For example, in computer vision, Mirjalili et al. (2020) proposed **PrivacyNet** to anonymize multiple face attributes (gender, age, ethnicity) via adversarial generative modeling.

For speech, Chen et al. (2024) recently proposed **MaSS**, which learns an encoder–decoder transformation to remove specified attributes from feature representations while preserving others and overall utility. MaSS employs an adversarial game between a suppressor network (trained to erase target attributes) and attribute classifiers (trained to detect residual traces), combined with contrastive losses to retain non-sensitive information. Their experiments across voice, image, and video domains showed that multiple attributes can be suppressed simultaneously without substantially harming downstream performance. However, their speech experiments were limited to the AudioMNIST dataset (short, fixed-length spoken digits under controlled conditions) and operated at the level of fixed-length embeddings (e.g., HuBERT representations Hsu et al., 2021) rather than directly on spectrograms or raw audio. This design involves training complex encoder–decoder models with adversarial and contrastive objectives, leaving an open gap for lightweight, post-hoc methods applicable to more natural, variable-length speech.

Our work addresses this gap by introducing an adversarial, inference-time approach for attribute-specific privacy control in speech. Rather than training a generative model to resynthesize obfuscated audio (Aloufi et al., 2019), we assume access to a pre-trained attribute classifier and directly manipulate input spectrograms to confuse specific attribute predictions. This design offers a lightweight, model-agnostic solution that avoids the complexity and computational overhead of generative methods, making it suitable for real-time or resource-constrained settings. Extending prior work beyond single-attribute masking, our method targets gender, age, or accent in real, variable-length speech using a multi-head classifier. By isolating the impact of each perturbation, we ensure minimal interference with non-target attributes—addressing a key limitation of existing approaches (Testa et al., 2023; Wang et al., 2024).

## 3. Proposed framework: Selective attribute masking (SAM)

The proposed framework consists mainly of a perturbation generation module that crafts small, targeted modifications to the input. These perturbations are designed to suppress the prediction of a specific attribute while preserving the rest of the speech content and non-target attribute predictions. As shown in Fig. 1, an untrusted third-party service may exploit machine learning models to infer sensitive attributes–such as gender, age, or accent – from users' voice data during cloud processing. SAM addresses this risk by applying targeted masking of selected attributes locally, before the data is transmitted, thereby preventing unwanted inference while preserving utility.

### 3.1. Model architecture

The proposed framework adopts a multi-task neural network architecture designed to jointly predict multiple speaker attributes from speech features. The central motivation for this design is to facilitate selective masking: by computing adversarial gradients with respect to individual output heads, the model can generate perturbations that suppress the prediction of a specific attribute while preserving the accuracy of others. This mechanism forms the foundation of the SAM method, enabling fine-grained, user-controllable privacy during inference.

As shown in Fig. 2, the model comprises three main components: input spectrograms, a shared feature extraction backbone, and multiple attribute-specific output heads. The input spectrograms are processed by a shared feature extractor composed of convolutional layers that capture local spectral characteristics, followed by recurrent layers that model temporal dependencies in the speech signal. The resulting shared representation is then forwarded to parallel classification heads, each responsible for predicting a specific attribute such as gender, age, or accent. This modular structure allows for independent gradient flow from each output head, which is essential for generating targeted adversarial perturbations during selective masking.

### 3.2. Selective attribute masking

Given a trained multi-head model for attribute prediction, our goal is to generate adversarial perturbations to input spectrograms that selectively impede the model's ability to recognize a particular attribute, while leaving all other attributes unaffected. Specifically, let $X$ denotes the original spectrogram (a matrix of time-frequency magnitudes), and the model output be a tuple of predictions $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m)$ corresponding to $m$ speaker attributes (e.g., age, gender, and country of origin). Given a target attribute index $k \in \{1, 2, \dots, m\}$ that the user wishes to mask (e.g., age), our goal is to generate a small perturbation $\Delta_k$ such that, when added to the input $X$, the predicted label $\hat{y}_k$ for the target attribute is altered, while the predictions $\hat{y}_j$ for all other attributes $j \neq k$ remain unaffected. Formally, for a target attribute index $k \in \{1, 2, \dots, m\}$, we define an adversarial example as:

$$\tilde{X}_k = X + \Delta_k, \quad \text{subject to} \quad \|\Delta_k\|_\infty \leq \epsilon,$$

such that the model satisfies:

- **Target attribute misclassification:** $f_k(\tilde{X}_k) \neq f_k(X)$, where $f_k(\cdot)$ denotes the model's prediction for attribute $k$.

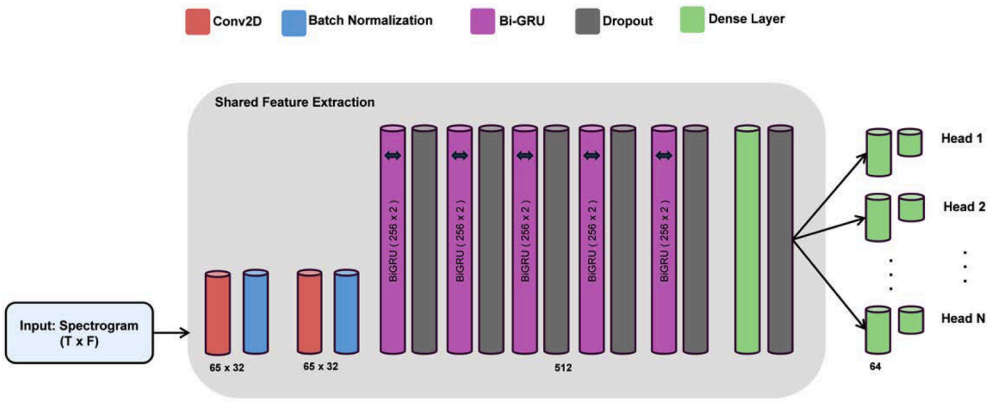



**Fig. 2.** Block diagram of the proposed multi-head model.

- **Non-target attribute preservation:** For all $j \in \{1, 2, \ldots, m\} \setminus \{k\}$, the predicted class remains unchanged, i.e., $f_j(\tilde{X}_k) = f_j(X)$.

In practice, exactly satisfying all these constraints is challenging due to shared representations among the attribute heads. Therefore, we approximate this selective behavior using gradient-based adversarial perturbations, guided by a custom composite loss that balances misclassification of the target attribute and preservation of the others. In this paper, this proposed composite loss function encourages misclassification of a specific target attribute while discouraging changes to the predictions of all non-target attributes, which can be defined as follows:

$$\mathcal{L}_{\text{composite}} = \mathcal{L}_k - \lambda \sum_{j \in \{1,2,\ldots,m\} \setminus \{k\}} \mathcal{L}_j \tag{1}$$

where $\mathcal{L}_k$ is the classification loss (e.g., cross-entropy) for the target attribute $k$, and $\mathcal{L}_j$ denotes the loss for each non-target attribute $j \in \{1, 2, \ldots, m\} \setminus \{k\}$. The hyperparameter $\lambda$ controls the trade-off between misclassifying the target and preserving the correctness of the remaining attributes. For all main experiments, we fix $\lambda = 1$ (see Section 5) to maintain balanced weighting between objectives. An ablation study on $\lambda$ is reported in Appendix A.1, confirming that larger values can further increase masking success but with less consistent utility preservation. The composite loss is specifically designed to enforce *selectivity* in the perturbation process. By maximizing the loss of the target attribute $\mathcal{L}_k$ while simultaneously minimizing the *average* loss of the non-target attributes $\mathcal{L}_j$ for all $j \neq k$, the optimizer is guided toward perturbations that induce *misclassification in only the desired head* while preserving the correctness of others. This stands in contrast to conventional adversarial attacks that optimize a single-head objective and may inadvertently disrupt multiple outputs in multi-task models.

The subtractive formulation of Eq. (1) creates a *gradient tension* between heads, where the perturbation must deceive the target classifier without degrading performance in others. This is particularly important in shared-representation architectures like ours, where gradients from different heads are entangled. By explicitly penalizing deviations in non-target predictions, the composite loss encourages *localized and purpose-driven* changes in the input–an essential requirement for privacy-preserving interventions that seek to alter one attribute while preserving utility across the rest.

To generate perturbations, we adapt two widely used attack algorithms: the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) and Projected Gradient Descent (PGD) (Madry et al., 2019)–as they offer simple and effective mechanisms to optimize our composite loss. FGSM is computationally efficient, applying a single-step perturbation:

$$\Delta_k = \epsilon \cdot \text{sign}\left(\nabla_X \mathcal{L}_{\text{composite}}\right),$$

**Algorithm 1** Generalized PGD with consistency-based composite loss for selective attribute masking.

**Input:** Input $X$; true labels $\{y_1, y_2, \ldots, y_m\}$; model $f$ with $m$ outputs; perturbation target $k \in \{1, \ldots, m\}$; number of steps $T$; step size $\alpha$; perturbation bound $\epsilon$; consistency weight $\lambda$; clipping bounds $[\text{clip}_{\min}, \text{clip}_{\max}]$
**Output:** Perturbed input $\tilde{X}_k$

**Initialize:** $\tilde{X}_k \leftarrow X$
**for** $t = 1$ to $T$ **do**
  **Forward pass:** $\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\} \leftarrow f(\tilde{X}_k)$
  **Compute losses:** $\mathcal{L}_i(\hat{y}_i, y_i)$ for $i = 1$ to $m$
  $\mathcal{L}_{\text{composite}} = \mathcal{L}_k - \lambda \sum_{j \in \{1,2,\ldots,m\} \setminus k} \mathcal{L}_j$
  **Update:** $\tilde{X}_k \leftarrow \tilde{X}_k + \alpha \cdot \text{sign}\left(\nabla_X \mathcal{L}_{\text{composite}}\right)$
  **Clip to** $\ell_\infty$ ball: $\tilde{X}_k \leftarrow \text{clip}(\tilde{X}_k, X - \epsilon, X + \epsilon)$
  **Clip to valid range:** $\tilde{X}_k \leftarrow \text{clip}(\tilde{X}_k, \text{clip}_{\min}, \text{clip}_{\max})$
**end for**
**Return** $\tilde{X}_k$

while PGD performs stronger, iterative updates with projection at each step, enabling more precise control over the perturbation:

$$\tilde{X}_k^{(t+1)} = \text{clip}_{\tilde{X}_k, \epsilon}\left(\tilde{X}_k^{(t)} + \alpha \cdot \text{sign}\left(\nabla_X \mathcal{L}_{\text{composite}}\right)\right),$$

where $\alpha$ is the step size and the projection ensures that the perturbation remains within the $\ell_\infty$ ball around the original input. Algorithm 1 summarizes the customized PGD procedure incorporating the consistency-based composite loss for SAM.

This formulation represents a subtle but important deviation from standard FGSM/PGD attacks, which typically use a single-head loss. By leveraging the multi-head structure and penalizing collateral changes, we enforce *selectivity* in the attack. While more complex regularization schemes could be used, we find that our framework with modest $\epsilon$ values and a consistency-weighted loss is often sufficient to achieve effective and targeted masking.

## 4. Experimental setup

### 4.1. Dataset and preprocessing

Our experiments use the English subset of the Mozilla Common Voice 17.0 corpus (Ardila et al., 2020). For comprehensive speaker coverage, we apply multiple filters. We select the top five English accent groups with over 40,000 samples, keeping utterances with valid accent, gender, and age labels, and at least two upvotes. We group speaker ages into three categories–*young*, *adult*, and *old*–based on similar mappings from (Tursunov et al., 2021). We removed German-accent samples due to gender imbalance (only male speakers). We use this curated subset because, to our knowledge, no other open corpus currently provides

utterance-level labels for age, gender, and accent at comparable scale; Common Voice is the only widely available source where all three attributes are available per clip via contributor metadata.

The dataset contains 22,212 utterances, annotated with accent (4 classes: United States English, England English, Indian Subcontinent English, Canadian English), gender (binary: male/female), and age group (3 classes: young, adult, old) for training the multi-head multi-class (MHMC) model. For the binary formulation (MHBC), we use 8000 balanced utterances with equal representation across attribute pairs. We split each dataset into 80 % training and 20 % testing, maintaining stratification across attributes. These splits apply to both MHMC and MHBC experiments. Audio samples are preprocessed into magnitude spectrograms using short-time Fourier transform (STFT) (Gabor, 1946) with frame length 256, hop size 128, and FFT size 256.

### 4.2. Models and training settings

To evaluate the proposed SAM framework, we first establish profiling models capable of accurately predicting speaker attributes. Two multi-task neural architectures were developed in this work and are used as profiling backbones to evaluate SAM: a Multi-Head Binary Classifier (MHBC) and a Multi-Head Multi-Class Classifier (MHMC). Both share a convolutional–recurrent backbone comprising two Conv2D layers followed by bidirectional GRUs, which process input spectrograms into a shared latent representation. This representation is passed to three task-specific dense heads for gender, accent, and age prediction. Each head includes a dense layer and an output layer with an activation function (sigmoid or softmax), depending on whether the binary (MHBC) or multi-class (MHMC) formulation is used. The MHBC serves as a simplified baseline to verify selective attribute masking under controlled conditions, whereas the MHMC reflects a more realistic multi-class configuration used for full SAM evaluation and privacy–utility analysis. Both backbones differ only in their output head configurations: MHBC employs binary heads for all attributes, whereas MHMC replaces the accent and age heads with multi-class layers.

Both models are trained using the Adam optimizer with a learning rate of 0.0001, a commonly adopted setting in audio and speech-related deep learning tasks, as it provided stable convergence in preliminary experiments. Gradient clipping (clip norm = 1.0) is applied to prevent exploding gradients and ensure training stability. To mitigate overfitting, we follow prior work on privacy-preserving speech models (Chandrinos et al., 2024) and apply dropout (rate = 0.5) together with L2 regularization. Training is performed for up to 50 epochs with a batch size of 32, values selected based on standard practice and confirmed to be computationally feasible for our setup. Loss weights are tuned empirically on the validation set to balance task contributions: gender loss weight is set to 1.0, and accent and age losses are weighted at 0.7 each, as this combination yielded the best trade-off across heads. Early stopping (patience = 15 epochs) and learning rate reduction (factor = 0.2, patience = 4) are applied based on validation accuracy, reflecting commonly used strategies for avoiding overfitting while ensuring efficient training. The final model is selected as the checkpoint with the lowest overall validation loss.

All training is conducted on an NVIDIA RTX 4000 SFF Ada Generation GPU with 20 GB of graphics memory. Consistent with the SAM framework, all adversarial experiments use a white-box threat model where the user has full access to model parameters and gradients, enabling gradient-based perturbation methods as described in Section 3.2.

### 4.3. Evaluation metrics

We evaluate our system using two categories of metrics. First, standard classification metrics (accuracy, precision, recall, and F1-score) are reported in Section 5.1 to validate the performance of the audio profiling models. These ensure that the underlying models are sufficiently accurate to serve as meaningful baselines for privacy-preserving experiments. Second, to assess the SAM framework itself, we employ task-specific privacy and utility metrics described below.

#### 4.3.1. Privacy metric – selective masking success rate

To evaluate the efficacy of SAM, we introduce a new metric called *Selective Masking Success Rate* (SMSR). This metric quantifies the proportion of test samples for which the perturbation successfully alters the target attribute, while all non-target attributes remain unchanged. Formally, for a given target attribute $k$, we consider an attack on a sample $X$ to be successful if:

$$f_k(\tilde{X}) \neq f_k(X) \quad \text{and} \quad f_j(\tilde{X}) = f_j(X) \quad \text{for all } j \neq k,$$

We define the overall success rate for attribute $k$ under perturbation budget $\epsilon$ as:

$$\text{SMSR}(k, \epsilon) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[ f_k(\tilde{X}) \neq f_k(X) \wedge \forall_{j \neq k}\, f_j(\tilde{X}) = f_j(X) \Big], \tag{2}$$

where $N$ is the number of test samples, and $\mathbf{1}[\cdot]$ is the indicator function that returns 1 if the condition holds and 0 otherwise.

#### 4.3.2. Utility evaluation – speech reconstruction and ASR

Ensuring privacy should not make speech unusable for its primary purpose (e.g., command recognition). A core part of our methodology is to evaluate the utility of adversarially perturbed audio. After applying spectrogram perturbation $\Delta_k$, we convert the modified spectrogram $\tilde{X}_k$ back into an audio waveform for utility evaluation. We use the Phase Gradient Heap Integration (PGHI) algorithm (Pruša & Søndergaard, 2016), a deterministic spectrogram inversion technique. PGHI estimates a phase matrix for $\tilde{X}_k$ and reconstructs time-domain audio without needing the original phase of $X$. PGHI exploits the relationship between phase gradients and magnitude in the time-frequency domain to recover a plausible phase that produces a time waveform when combined with the perturbed magnitude. We prefer PGHI over iterative Griffin-Lim reconstructions for efficiency and consistency (Holighaus et al., 2019). The result is an adversarial audio sample $\tilde{s}(t)$ corresponding to the perturbed spectrogram $\tilde{X}$. We evaluate the intelligibility and utility of $\tilde{s}(t)$ by transcribing it with OpenAI's Whisper model (Radford et al., 2023) (Medium edition). We compute the Word Error Rate (WER) between the ASR transcription of the perturbed audio and the reference transcription. The WER measures how much speech content was distorted by our perturbation process. A low WER indicates preserved speech intelligibility and linguistic information. We conduct subjective checks on random samples to ensure imperceptible perturbations: the noise is not noticeable at chosen $\epsilon$ levels, and listeners cannot detect differences in speaker attributes (Wang et al., 2024). All experiments use careful calibration of $\epsilon$ to balance privacy-utility trade-off: we seek minimal perturbation that achieves high selective masking success while maintaining utility.

Our methodology integrates a multi-task attribute model with adversarial attack techniques for selective voice attribute masking. By measuring privacy gains and utility impact, we provide comprehensive evaluation. Together, these evaluations provide a comprehensive framework for assessing the privacy–utility trade-off achieved by SAM, as reported in the following section.

## 5. Result and discussion

This section presents a comprehensive evaluation of the SAM framework. We investigate its effectiveness using two multi-head model variants: MHBC and MHMC. Our analysis focuses on two key dimensions: *privacy*, measured by the selective masking success rate across varying adversarial perturbation strengths ($\epsilon$), and *utility*, assessed via WER computed after reconstructing perturbed audio signals. The central objective is to determine whether adversarial perturbations can reliably suppress inference of a targeted attribute while preserving the accuracy of non-targeted attributes and maintaining utility.

**Table 1**
Performance metrics of MHBC model for gender, accent, and age classification.

| Task | Class | Precision | Recall | F1 | Accuracy (%) |
|---|---|---|---|---|---|
| Gender (Binary) | Female | 0.955 | 0.983 | 0.969 | 96.8 |
| | Male | 0.982 | 0.954 | 0.968 | |
| Accent (Binary) | British | 0.900 | 0.866 | 0.883 | 88.5 |
| | Canadian | 0.871 | 0.904 | 0.887 | |
| Age (Binary) | Young | 0.918 | 0.966 | 0.942 | 94.0 |
| | Old | 0.964 | 0.914 | 0.938 | |

**Table 2**
Performance metrics of MHMC model for gender, accent, and age classification.

| Task | Class | Precision | Recall | F1 | Accuracy (%) |
|---|---|---|---|---|---|
| Gender (Binary) | Female | 0.954 | 0.977 | 0.965 | 96.5 |
| | Male | 0.977 | 0.954 | 0.965 | |
| Accent (4-Class) | British | 0.720 | 0.703 | 0.712 | 71.3 |
| | Canadian | 0.752 | 0.735 | 0.743 | |
| | US | 0.615 | 0.655 | 0.634 | |
| | IS | 0.799 | 0.777 | 0.788 | |
| Age (3-Class) | Young | 0.640 | 0.821 | 0.719 | 73.5 |
| | Adult | 0.730 | 0.598 | 0.657 | |
| | Old | 0.920 | 0.800 | 0.855 | |

### 5.1. Experimental results for audio profiling models

Tables 1 and 2 report detailed per-class performance for MHBC and MHMC, evaluated using standard classification metrics (accuracy, precision, recall, F1-score). The MHBC achieved strong accuracies across all attributes (96.8 % for gender, 88.5 % for accent, 94.0 % for age), with relatively balanced precision–recall trade-offs across classes. In contrast, the MHMC maintained stable performance for gender (96.5 %) but showed reduced accuracy for accent (71.3 %) and age (73.5 %). Performance drops were most notable for the U.S. accent and the adult age class, where inter-class similarity likely caused confusion. These results indicate that MHBC achieves stronger baseline accuracies overall, while MHMC provides a more realistic multi-class formulation. Accordingly, both models are carried forward as profiling backbones for the SAM experiments in the subsequent section.

### 5.2. Selective masking success rate

We report the success rate of adversarial attacks targeted at each individual attribute (gender, accent, age), under the condition that the other two attributes remain correctly classified. Figures and tables illustrate success rates as a function of $\epsilon \in [0.0, 0.2]$ for both MHBC and MHMC models. This range is consistent with prior adversarial robustness literature, where small $\ell_\infty$ perturbation budgets are commonly used to ensure imperceptibility and maintain semantic fidelity (Goodfellow et al., 2015; Madry et al., 2019; Siedel2024, 2024), and is chosen here to minimize degradation in utility, which we quantify using WER. Unless otherwise specified, we use a maximum perturbation budget of $\epsilon = 0.2$, step size $\alpha = 0.01$, number of PGD steps $T = 15$, and clipping bounds $[\text{clip}_{\min}, \text{clip}_{\max}] = [-3, 3]$ to ensure that adversarial examples remain within the typical dynamic range of spectrogram inputs, additionally we set the number of attributes to $m = 3$ and use $\lambda = 1$ in the composite loss function (Eq. (1)), ensuring balanced emphasis between the target and averaged non-target loss terms during perturbation generation.

For the MHBC model, the highest success rate is observed in the age masking task (up to 74.5 % at $\epsilon = 0.2$ using PGD), followed by gender (59.4 %) and accent (54.6 %). MHMC results follow a similar trend, though success rates are slightly lower for age (62.3 %), followed by gender (47.6 %) and higher for accent (58.2 %). These outcomes are significantly better than random perturbation baselines, which yield success rates below 20 % for all attributes, confirming the selectivity and effectiveness of our framework. Tables 3 and 4 summarize the results across the full $\epsilon$ range.

We further compare the performance of our custom FGSM and custom PGD implementations. As expected, PGD consistently outperforms FGSM across all attributes and both model architectures due to its iterative refinement. For MHBC, at $\epsilon = 0.05$, FGSM achieves 46.2 % success for age, 18.5 % for gender, and 32.8 % for accent, while PGD yields 49.3 %, 31 %, and 31.7 % respectively. Similarly, for MHMC at the same $\epsilon$, FGSM achieves 47.53 % for age, 14.06 % for gender, and 42.78 % for accent, compared to PGD values of 48.59 %, 16.36 %, and 45.41 %. Despite being a single-step method, FGSM remains competitive at lower perturbation levels and offers a computationally efficient alternative to PGD.

### 5.3. WER-based utility analysis

To evaluate the utility of perturbed speech, we measure WER using a Whisper-based transcription system. Our baseline (unperturbed) validation set yields a WER of 8.2 %, and reconstruction from spectrogram alone slightly increases this to 8.56 % for MHBC, confirming that the reconstruction pipeline preserves intelligibility. For custom FGSM (See Fig. 3a), WER increases gradually from 10.4 % at $\epsilon = 0.01$ to 12.17 % at $\epsilon = 0.2$ (rounded values). For PGD (See Fig. 3b), the increase is even more controlled, with WER ranging from 10.4 % to 10.79 %.

Notably, the average WER increase relative to reconstructed audio remained under 3 % absolute for both FGSM and PGD across all tested $\epsilon$ values. PGD results showed remarkable stability, with WER peaking around $\epsilon = 0.125$ and then plateauing. These results indicate that perturbations were applied effectively while distortion remained bounded at moderate perturbation strengths.

Importantly, even as WER stabilizes, the success rate of attribute masking continues to increase across all heads – from 63.1 % to 74.5 % for age, 45.3 % to 59.4 % for gender, and 49.4 % to 54.6 % for accent, as $\epsilon$ increases from 0.125 to 0.2. This indicates that PGD effectively converges to regions of the input space that achieve high privacy with minimal utility degradation. These findings demonstrate that SAM preserves ASR performance while substantially enhancing privacy. Taken together, this positions our framework as a practical solution for balancing privacy and utility in voice-based systems.

A slight increase in WER is observed when moving from MHBC to MHMC, even before adversarial perturbation is applied. Specifically, the original WER for MHBC was 8.2 % (1600 samples), whereas MHMC yielded 8.53 % (4443 samples). Similarly, reconstruction WER increased from 8.56 % to 9.17 %. This shift is expected, given the broader and more diverse input distribution in the MHMC dataset, which includes acoustically varied samples. A key observation emerges when comparing the impact of adversarial attacks across models. For MHBC, PGD consistently introduced less transcription distortion than FGSM–e.g., at $\epsilon = 0.2$, PGD resulted in a WER of 10.79 %, compared to FGSM's 12.17 %. However, for MHMC (See Fig. 4, the difference between FGSM and PGD is less pronounced: WERs at $\epsilon = 0.2$ are 13.93 % (FGSM) and 14.09 % (PGD), a gap of just 0.16 % absolute. Both values remain within 6 % absolute error from the unperturbed reconstruction, underscoring the robustness of the perturbation design in preserving ASR utility.

This convergence in WER between FGSM and PGD under the MHMC setting may be attributed to the model's increased output complexity. Unlike MHBC, which uses binary classification heads, MHMC handles multi-class outputs for each attribute, resulting in larger softmax spaces and potentially more diffused gradient signals. This complexity could limit the degree to which iterative refinement (as in PGD) can further optimize perturbations without incurring additional utility cost. Nonetheless, despite their similar WER outcomes, PGD significantly outperforms FGSM in terms of selective masking success. At $\epsilon = 0.2$, FGSM achieves success rates of 40.0 % (age), 19.4 % (gender), and 38.4 %

**Table 3**
Selective masking success rates (%) for MHBC using custom FGSM and PGD, compared against random baselines.

| Method | Attribute | Epsilon ($\epsilon$) | | | | | | | | Baseline |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.01 | 0.03 | 0.05 | 0.10 | 0.125 | 0.15 | 0.175 | 0.2 | |
| FGSM | Age | 13.81 | 32.94 | 46.19 | 53.86 | 54.00 | 52.38 | 51.50 | 51.00 | 05.14 |
| | Gender | 10.06 | 15.06 | 18.50 | 25.38 | 27.31 | 29.00 | 29.56 | 30.44 | 02.65 |
| | Accent | 17.63 | 30.18 | 32.75 | 30.38 | 30.86 | 31.19 | 31.13 | 30.56 | 10.47 |
| PGD | Age | 18.69 | 50.25 | 49.25 | 56.13 | 63.13 | 68.13 | 73.13 | **74.50** | 05.14 |
| | Gender | 14.06 | 32.13 | 31.00 | 39.25 | 45.31 | 50.19 | 56.19 | **59.38** | 02.65 |
| | Accent | 22.50 | 37.81 | 31.69 | 43.19 | 49.38 | 52.94 | 53.13 | **54.56** | 10.47 |

**Table 4**
Selective masking success rates (%) for MHMC using custom FGSM and PGD, compared against random baselines.

| Method | Attr | Epsilon $\epsilon$ | | | | | | | | Baseline |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0.01 | 0.03 | 0.05 | 0.10 | 0.125 | 0.15 | 0.175 | 0.20 | |
| FGSM | Age | 26.24 | 43.03 | 47.53 | 45.05 | 43.55 | 41.93 | 41.16 | 40.01 | 18.28 |
| | Gender | 5.69 | 11.16 | 14.06 | 17.62 | 18.20 | 18.92 | 19.06 | 19.35 | 1.81 |
| | Accent | 26.82 | 38.44 | 42.78 | 40.78 | 39.50 | 38.91 | 38.66 | 38.35 | 20.36 |
| PGD | Age | 34.32 | 52.75 | 48.59 | 54.69 | 59.73 | 59.62 | 59.64 | **62.32** | 18.28 |
| | Gender | 11.41 | 19.26 | 16.36 | 36.43 | 41.61 | 44.54 | 47.19 | **47.64** | 1.81 |
| | Accent | 34.95 | 48.57 | 45.41 | 52.26 | 53.02 | 54.60 | 55.61 | **58.20** | 20.36 |

(accent), whereas PGD yields 62.3 %, 47.6 %, and 58.2 % respectively–showing consistent and substantial gains across all attributes. This reinforces PGD's superiority in privacy preservation, even under utility-constrained settings.

In summary, although MHMC experiences slightly higher transcription degradation than MHBC, it maintains acceptable WER under both FGSM and PGD. More importantly, PGD offers a more effective privacy–utility trade-off, achieving high masking success with minimal added distortion (See Fig. 5a and b). These results further validate our framework's generalizability across model architectures and input distributions, and are further illustrated by joint plots of average WER and SMSR as a function of perturbation strength $\epsilon$, which highlight key trends across models, perturbation budgets, and target attributes.

### 5.4. Comparative analysis with baselines

#### 5.4.1. Comparison with random baseline

While a direct comparable study for evaluating multi-head models under selective perturbation constraints could not be found, we establish a conservative baseline for success rates under random perturbations. This baseline corresponds to the joint probability of (i) the target attribute being misclassified at its natural error rate, and (ii) all non-target attributes being preserved at their clean accuracies–an optimistic assumption that likely overestimates non-target preservation under adversarial conditions. Following prior work that reports chance-level references for comparison (Aloufi et al., 2020; Wu et al., 2024), we adopt a random guessing baseline .

Using the MHBC model's test accuracies (96.81 % gender, 88.50 % accent, 94.00 % age), the expected random success rates are computed as the product of the target attribute's natural error rate $(1 - \mathrm{Acc}_k)$ and the clean accuracies of the two non-target attributes, i.e.,

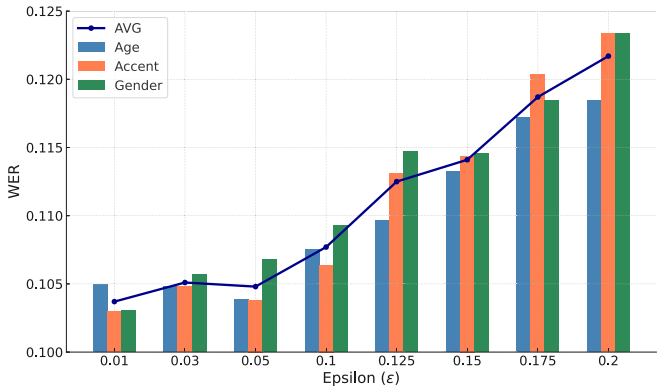
$$P_{\text{rand}}(k) = (1 - \mathrm{Acc}_k) \times \prod_{j \neq k} \mathrm{Acc}_j.$$

This yields 2.65 % (gender), 10.47 % (accent), and 5.14 % (age), which serve as optimistic upper bounds for success under purely random perturbations, as they assume perfect preservation of non-target attributes. In contrast, the proposed method achieves significantly higher success rates: 59.4 % (gender), 54.6 % (accent), and 74.5 % (age)–representing improvements of 22×, 5×, and 14× over random chance, respectively. These results provide a conservative lower bound on the attack's effectiveness, as adversarial perturbations typically degrade non-target performance in practice (Mahmood & Elhamifar, 2024), which would further reduce the random baseline. The substantial gaps between the observed and theoretical success rates confirm the precision of the targeted perturbations and, from a privacy perspective, demonstrate that the framework can reliably mask the chosen sensitive attribute while preserving overall speech utility to a large extent. Additional results evaluating SMSR using model-predicted labels–rather than ground truth—are provided in Fig. 6, further supporting the framework's applicability in real-world deployment scenarios where true attribute labels may not be accessible.

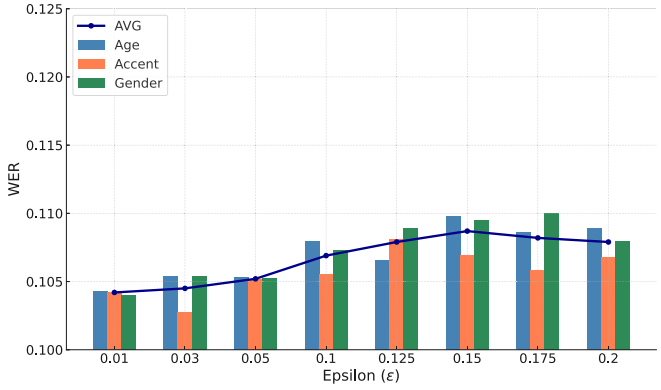
Further evaluation of selective masking performance on the MHMC model shows test accuracies of 96.53 % (gender), 71.28 % (accent), and 73.46 % (age). Under the same random baseline formulation, the expected success rates are 1.81 % (gender), 20.36 % (accent), and 18.26 % (age). SAM significantly outperforms these baselines, achieving selective masking success rates of 47.64 % (gender), 58.2 % (accent), and 62.32 % (age), corresponding to improvements of 26×, 2.9×, and 3.4×, respectively. Notably, the gender masking success rate exceeds its random baseline by more than an order of magnitude, despite the model's high clean accuracy for gender. Interestingly, however, MHMC does not consistently outperform the simpler MHBC variant: MHBC achieves higher absolute success rates in most attributes, especially for age (74.5 % vs. 62.32 %). In contrast, MHMC occasionally demonstrates stronger relative improvements over its random baselines, particularly for gender, where its natural error rate is lower. These differences may be explained by class imbalance, increased label sparsity in the multi-class setting, or overfitting challenges associated with higher output dimensionality. Collectively, these results validate the scalability of SAM to models with greater architectural complexity while also showing that MHBC remains a robust and efficient reference baseline under limited data or high-variance conditions.

#### 5.4.2. Comparison with single head classifier

This subsection compares the selective masking performance of the MHMC architecture with a baseline constructed from three independently trained single-head classifiers (SHC). The SHC models achieve strong standalone accuracies of 98.2 % (gender), 90.1 % (accent), and 94.2 % (age), reflecting the benefit of dedicating full model capacity

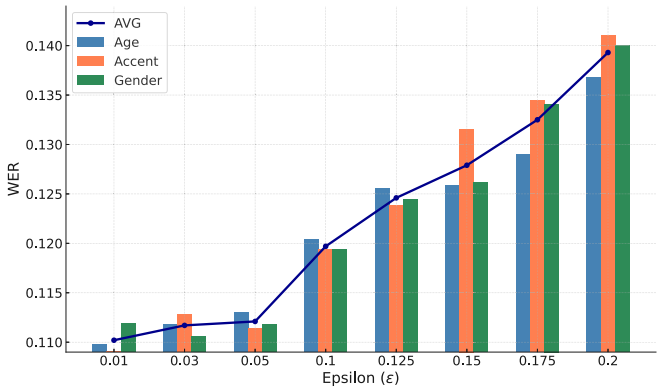



(a) FGSM (MHBC)

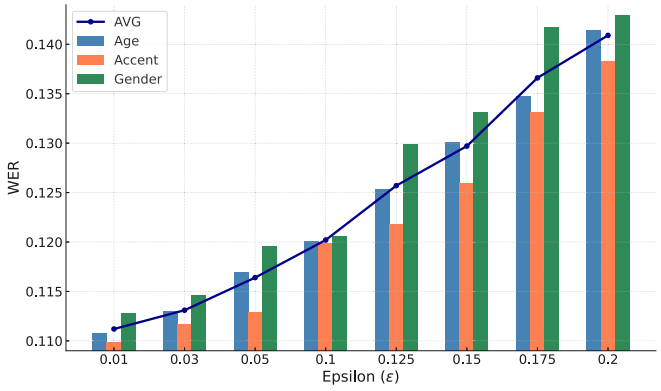



(b) PGD (MHBC)

**Fig. 3.** WER analysis for the MHBC model under different perturbation methods and strengths. (a) and (b) show per-attribute and average WER across $\epsilon$ for FGSM and PGD respectively.




(a) FGSM (MHMC)




(b) PGD (MHMC)

**Fig. 4.** WER analysis for the MHMC model under different perturbation methods and strengths. (a) and (b) show per-attribute and average WER across $\epsilon$ for FGSM and PGD respectively.

to each attribute. While earlier sections included results for MHBC and MHMC under both FGSM and PGD attacks, this section compares SHC only against MHMC using customized PGD-based perturbations, as MHMC is the final model selected for detailed evaluation and PGD consistently demonstrated superior masking performance in prior experiments. All results reported in this subsection–both selective masking (SMSR) and utility (WER)–were obtained using 4443 utterances from the validation split, corresponding to the 20 % of the 22,212-utterance dataset. This subset was used consistently across all SHC and MHMC evaluations to ensure comparability.

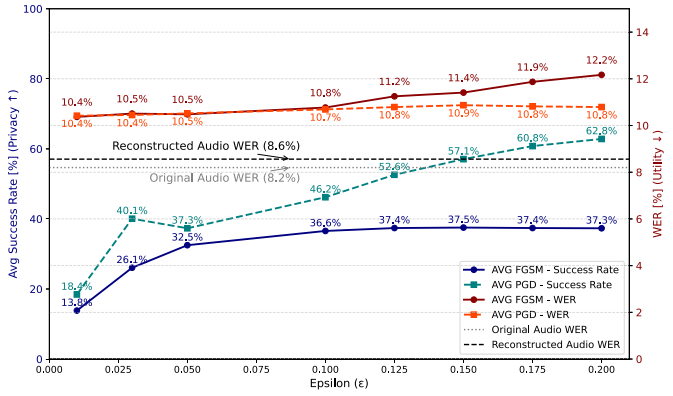
To enable a comparable selective masking experiment for SHC, perturbations were generated for each masking task (e.g., masking gender) by optimizing the same composite loss function defined in Eq. (1). In this case, $\mathcal{L}_k$ corresponds to the loss from the SHC model for the target attribute $k$, while the consistency terms $\mathcal{L}_j$ were computed using the predictions from the two other independently trained SHC models for the non-target attributes $j \neq k$. In contrast, the MHMC configuration computes all loss terms within a single multi-head network, where predictions for all attributes are derived from a shared representation.

In both setups, perturbations were constrained under the same $\ell_\infty$ budget ($\epsilon$) and evaluated using the SMSR. The results demonstrate that the SHC architecture consistently achieves higher SMSR across most perturbation strengths:

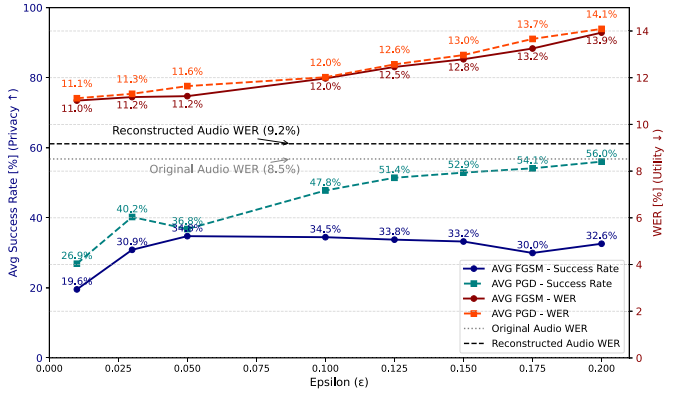
- **Accent and Gender masking**: SHC outperforms MHMC at all perturbation levels. At $\epsilon = 0.2$, SHC achieves SMSR of 98.04 % for accent and 97.93 % for gender, compared to 58.20 % and 47.64 % respectively for MHMC.
- **Age masking**: MHMC achieves marginally higher SMSR at very small perturbations ($\epsilon \leq 0.05$), reflecting greater initial sensitivity, but SHC overtakes beyond $\epsilon = 0.1$, achieving 92.57 % at $\epsilon = 0.2$ compared to 62.32 % for MHMC.

These findings highlight that SHC provides greater flexibility for selective masking, enabling nearly complete suppression of target attributes (See Fig. 7) at moderate perturbation levels–a desirable property when maximizing masking effectiveness is the primary objective. However, this capability comes without structural constraints to protect non-target attributes and requires maintaining separate models for each task. By contrast, MHMC's shared encoder introduces natural constraints on perturbations: perturbing one attribute while preserving others is inherently more difficult, which limits masking effectiveness but promotes consistency and utility preservation for non-target predictions. This constrained behavior aligns with MHMC's design objective of providing an integrated and balanced privacy-preserving solution across multiple attributes. Moreover, while SHC achieves higher masking success rates, it comes at the cost of training and deploying three independent models, incurring greater computational and storage overhead. MHMC, on the other hand, offers a unified architecture capable of multi-attribute prediction and selective masking within a single framework, trading off some masking success for better scalability, resource efficiency, and utility retention.

To complement these privacy-focused results, we next examine utility. The WER resulting from SHC and MHMC perturbations was compared across increasing $\ell_\infty$ budgets. As shown in Fig. 8, MHMC consistently produced lower WERs than SHC across all masking tasks (accent, age, and gender), particularly at higher $\epsilon$ values. For example,




(a) MHBC: WER vs Privacy




(b) MHMC: WER vs Privacy

**Fig. 5.** Joint analysis of privacy (selective masking success rate) and utility (WER) across perturbation strength $\epsilon$ for MHBC and MHMC models. PGD consistently achieves stronger attribute masking at comparable or lower utility cost than FGSM.

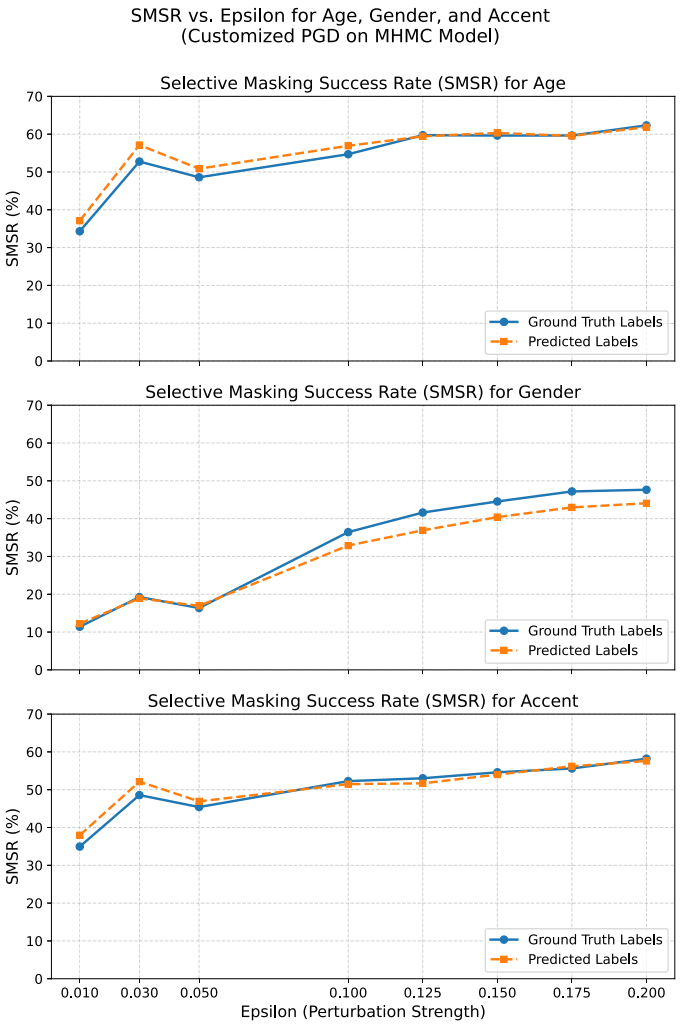



**Fig. 6.** SMSR for age, gender, and accent under PGD perturbations using predicted labels.

at $\epsilon = 0.2$, SHC exhibited WERs of 11.89 % (accent), 11.56 % (age), and 11.87 % (gender), whereas MHMC retained lower values of 10.68 %, 10.89 %, and 10.80 % respectively.

These results support the interpretation that MHMC's shared representation enforces perturbation constraints that not only limit aggressive masking but also better preserve overall speech intelligibility. In contrast, SHC's independent task-specific perturbations, while more successful at attribute masking, cause greater degradation in downstream utility. The average WER trend lines across perturbation budgets confirm that MHMC achieves a better privacy-utility balance overall, aligning with its design goals.

While the average WER shows a rising trend across the $\epsilon \in [0.01, 0.2]$ range, the relationship is not strictly monotonic. Local fluctuations were observed–for instance, WER for SHC age peaked at $\epsilon = 0.175$ before decreasing at $\epsilon = 0.2$, and MHMC gender WER followed a similar pattern. These non-linear trends may occur when stronger perturbations, instead of targeting the most ASR-relevant regions of the spectrogram, either miss them or diffuse across less important areas. As a result, the expected increase in transcription error does not always follow a linear path. However, when the perturbation budget is extended beyond this range (e.g., $\epsilon = 0.4$, 0.9), the WER increases significantly and consistently, reinforcing the notion that larger distortions reliably degrade speech intelligibility.

Taken together, these comparisons show that while SHC achieves stronger attribute suppression, MHMC provides a more balanced trade-off between privacy and utility, validating SAM as a practical and scalable framework for multi-attribute voice privacy.

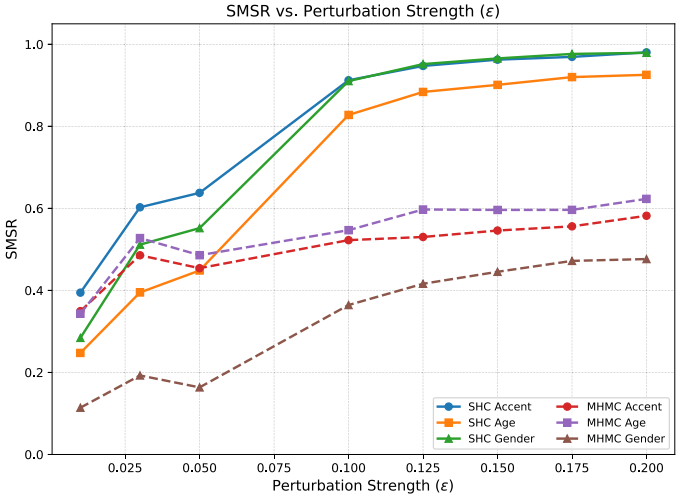



**Fig. 7.** Comparison of SMSR for MHMC and SHC architectures across varying $\epsilon$, for accent, age, and gender attributes.

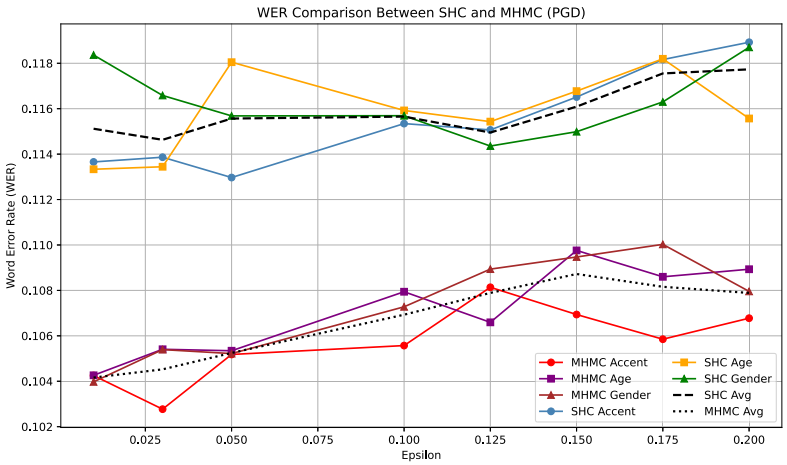



**Fig. 8.** WER comparison of MHMC and SHC architectures across varying $\epsilon$, for accent, age, and gender masking tasks.

## 6. Conclusion

This paper presents Selective Attribute Masking (SAM), a framework designed to enhance voice privacy not through complete anonymization, but by allowing users to selectively conceal specific attributes–such as age, gender, or accent–on demand through adversarial perturbations. SAM redefines privacy in speech as a selective, user-driven goal, where only sensitive inferences are obstructed while maintaining downstream functionalities like ASR. Our findings reveal that this framework is not only technically viable but also highly effective: SAM achieves attribute masking rates of up to 74.5 % with minimal impact on transcription quality when evaluated using multi-head classifier architectures. Additionally, a comparison with SHC baselines demonstrated that while SHC models achieved higher masking success–up to 98.04 %–this came at a notable cost to utility and scalability. In contrast, SAM's integration with a unified MHMC architecture preserved utility more effectively while still delivering meaningful masking performance, highlighting its practical suitability for real-time, post-hoc privacy control.

In future work, we plan to broaden the scope of SAM's validation to include a more diverse range of speaker attributes, such as emotional states, health cues, and environmental contexts, once suitable annotated datasets become available. These attributes introduce new privacy risks, and the ability to selectively mask them could support more adaptive and user-controlled speech interfaces. We also intend to evaluate SAM under a black-box threat model, where the adversary has limited or no access to model parameters or gradients. This setting better reflects realistic deployment scenarios such as third-party inference services and poses new challenges for crafting effective perturbations. SAM represents a first step toward aligning AI voice technologies with real-world privacy needs, where control lies not with the system, but with the speaker.

## CRediT authorship contribution statement

**Anil Pudasaini:** Conceptualization, Data curation, Methodology, Software, Writing – original draft, Writing – review & editing, Visualization; **Muna Al-Hawawreh:** Writing – review & editing, Supervision, Validation; **Mohamed Reda Bouadjenek:** Project administration, Supervision, Methodology, Validation; **Hakim Hacid:** Funding acquisition, Conceptualization; **Sunil Aryal:** Writing – review & editing, Supervision, Validation.

## Data availability

Data will be made available on request.

## Declaration of competing interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mohamed Reda Bouadjenek reports financial support was provided by Technology Innovation Institute. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

This research is supported by the Technology Innovation Institute, UAE under the research contract number TII/DSRC/2022/3143.

## Appendix A. Additional experiments

### A.1. Ablation study on $\lambda$

As discussed in Section 3.2, the hyperparameter $\lambda$ in the composite loss balances target suppression against preservation of non-target attributes. While all main experiments fixed $\lambda = 1$ for balanced weighting, here we report an ablation study to examine the effect of larger values at $\epsilon = 0.2$.

Fig. A.9 shows SMSR for age, accent, and gender under varying $\lambda \in \{1, 5, 10, 15, 20\}$. The results indicate that increasing $\lambda$ improves masking effectiveness for age and accent, with age SMSR rising from 62.3 % at $\lambda = 1$ to 80.1 % at $\lambda = 20$, and accent improving from 58.2 % to 86.5 % over the same range. Gender shows only modest gains, increasing from 47.6 % to about 57.5 %, and then plateauing beyond $\lambda = 10$. These findings show that higher $\lambda$ values substantially improve SMSR for age and accent, while gains for gender plateau beyond moderate values.

Table A.5 reports the corresponding Word Error Rate (WER) values for the same settings. WER remains relatively stable across $\lambda \in \{1, 5, 10, 15\}$, fluctuating within a narrow 1–1.5 % band. Notably, age WER shows a non-monotonic trend–peaking at $\lambda = 5$ (15.2 %) before decreasing again at $\lambda = 15$ (14.2 %). Accent and gender WERs increase slightly with larger $\lambda$, but remain below 15.1 %. This indicates that while stronger consistency weighting improves masking success, utility preservation does not follow a strictly monotonic trajectory.

Overall, the ablation confirms that larger $\lambda$ values can further increase masking success, but improvements come with diminishing returns and sometimes inconsistent utility preservation. This supports the choice of $\lambda = 1$ in the main experiments, which provides a balanced privacy–utility trade-off.

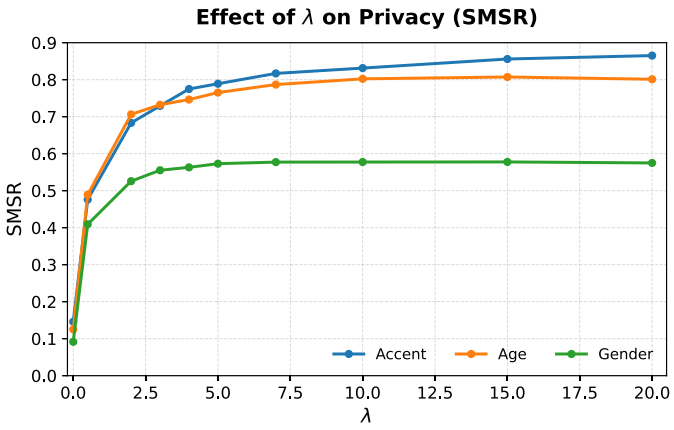



**Fig. A.9.** Effect of $\lambda$ on SMSR at $\epsilon = 0.2$. Larger values enhance masking effectiveness, particularly for age and accent.

**Table A.5**
WER results at $\epsilon = 0.2$ across different $\lambda$ values (lower is better).

| $\lambda$ | Accent WER (%) | Age WER (%) | Gender WER (%) |
|---|---|---|---|
| 1 | 13.82 | 14.14 | 14.30 |
| 5 | 14.25 | 15.22 | 14.44 |
| 10 | 14.88 | 14.46 | 14.82 |
| 15 | 15.06 | 14.20 | 14.63 |

## References


Ali, L., Zhu, C., Zhou, M., & Liu, Y. (2019). Early diagnosis of Parkinson's disease from multiple voice recordings by simultaneous sample and feature selection. *Expert Systems with Applications*, *137*, 22–28. https://doi.org/10.1016/j.eswa.2019.06.052

Aloufi, R., Haddadi, H., & Boyle, D. (2019). Emotionless: Privacy-preserving speech analysis for voice assistants. http://arxiv.org/abs/1908.03632.

Aloufi, R., Haddadi, H., & Boyle, D. (2020). Privacy-preserving voice analysis via disentangled representations. In *Proceedings of the 2020 ACM SIGSAC conference on cloud computing security workshop* CCSW'20 (p. 1–14). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3411495.3421355

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th conference on language resources and evaluation (LREC 2020)* (pp. 4211–4215).

Cai, Z., Xinyuan, H. L., Garg, A., Garcéa-Perera, L. P., Duh, K., Khudanpur, S., Andrews, N., & Wiesner, M. (2024). Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization. https://arxiv.org/abs/2409.03655.

Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., & Ponti, M. A. (2023). YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. https://arxiv.org/abs/2112.02418.

Chandrinos, N., Loi, I., Zachos, P., Symeonidis, I., Spiliotis, A., Panou, M., & Moustakas, K. (2024). Effectiveness of l2 regularization in privacy-preserving machine learning. https://arxiv.org/abs/2412.01541.

Chen, M., Lu, L., Yu, J., Chen, Y., Ba, Z., Lin, F., & Ren, K. (2022). Privacy-utility balanced voice de-identification using adversarial examples. https://arxiv.org/abs/2211.05446.

Chen, Y., Chen, C.-F., Hsu, H., Hu, S., Pistoia, M., & Abdelzaher, T. (2024). MaSS: Multi-attribute selective suppression for utility-preserving data transformation from an information-theoretic perspective. https://arxiv.org/abs/2405.14981.

Chouchane, O., Panariello, M., Zari, O., Kerenciler, I., Chihaoui, I., Todisco, M., & Önen, M. (2023). Differentially private adversarial auto-encoder to protect gender in voice biometrics. https://arxiv.org/abs/2307.02135.

Gabor, D. (1946). Theory of communication. *Journal of the Institution of Electrical Engineers - Part III: Radio and Communication Engineering*, *93*(26), 429–457.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572.

Holighaus, N., Koliander, G., Abreu, L. D., & Průša, Z. (2019). Non-iterative phaseless reconstruction from wavelet transform magnitude. In *Proceedings of the 22nd international conference on digital audio effects (DAFx-19)* (pp. 1–8). Birmingham, UK. https://ltfat.org/notes/ltfatnote055.pdf.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. https://arxiv.org/abs/2106.07447.

Jaiswal, M., & Provost, E. M. (2019). Privacy enhanced multimodal neural representations for emotion recognition. https://arxiv.org/abs/1910.13212.

Kröger, J. L., Lutz, O. H.-M., & Raschke, P. (2020). Privacy implications of voice and speech analysis – information disclosure by inference. In *Privacy and identity management. data for better living: AI and privacy: 14th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2.2 international summer school, Windisch, Switzerland, August 19–23, 2019, revised selected papers* IFIP Advances in Information and Communication Technology (pp. 242–258). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-42504-3_16

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks. https://arxiv.org/abs/1706.06083.

Mahmood, H., & Elhamifar, E. (2024). Semantic–aware multi–label adversarial attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 24251–24262). https://doi.org/10.1109/CVPR52733.2024.02289

Mirjalili, V., Raschka, S., & Ross, A. (2020). Privacynet: Semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing : A Publication of the IEEE Signal Processing Society*, *PP*. https://doi.org/10.1109/TIP.2020.3024026

Mohammed, H. M. A., Omeroglu, A. N., & Oral, E. A. (2023). MMHFNet: Multi-modal and multi-layer hybrid fusion network for voice pathology detection. *Expert Systems with Applications*, *223*, 119790. https://www.sciencedirect.com/science/article/pii/S0957417423002919.

Muthalagu, R., Malik, J., & Pawar, P. M. (2025). Detection and prevention of evasion attacks on machine learning models. *Expert Systems with Applications*, *266*, 126044. https://doi.org/10.1016/j.eswa.2024.126044

Patino, J., Tomashenko, N., Todisco, M., Nautsch, A., & Evans, N. (2021). Speaker anonymisation using the McAdams coefficient. In *Proc. interspeech 2021* (pp. 1099–1103). https://doi.org/10.21437/Interspeech.2021-1070

Pruša, Z., & Søndergaard, P. L. (2016). Real-time spectrogram inversion using phase gradient heap integration. In *Proc. int. conf. digital audio effects (DAFx-16)* (pp. 17–21).

Pudasaini, A., Al-Hawawreh, M., Bouadjenek, M. R., Hacid, H., & Aryal, S. (2025). A comprehensive study of audio profiling: Methods, applications, challenges, and future directions. *Neurocomputing*, (p. 130334). https://doi.org/https://doi.org/10.1016/j.neucom.2025.130334

Rabhi, M., Bakiras, S., & Di Pietro, R. (2024). Audio-deepfake detection: Adversarial attacks and countermeasures. *Expert Systems with Applications*, *250*, 123941. https://doi.org/10.1016/j.eswa.2024.123941

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518). PMLR.

Siedel2024, (2024). A practical approach to evaluating the adversarial distance for machine learning classifiers (*vol. Volume 11: Safety Engineering, Risk and Reliability Analysis; Research Posters*). ASME International Mechanical Engineering Congress and Exposition. ASME International. https://doi.org/10.1115/IMECE2024-145280

Singh, R. (2019). Reconstruction of the human persona in 3D from voice, and its reverse. In R. Singh (Ed.), *Profiling humans from their voice* (pp. 325–363). Singapore: Springer. https://doi.org/10.1007/978-981-13-8403-5_9

Stoidis, D., & Cavallaro, A. (2022). Generating gender-ambiguous voices for privacy-preserving speech recognition. In *Interspeech 2022* (pp. 4237–4241). ISCA. https://doi.org/10.21437/Interspeech.2022-11322

Testa, B., Xiao, Y., Sharma, H., Gump, A., & Salekin, A. (2023). Privacy against real-time speech emotion detection via acoustic adversarial evasion of machine learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *7*(3), 1–30. https://doi.org/10.1145/3610887

Tomashenko, N., Miao, X., Champion, P., Meyer, S., Wang, X., Vincent, E., Panariello, M., Evans, N., Yamagishi, J., & Todisco, M. (2024). The VoicePrivacy 2024 Challenge Evaluation Plan. arXiv:2404.02677 [cs, eess] http://arxiv.org/abs/2404.02677.

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). The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, *74*(C). https://doi.org/10.1016/j.csl.2022.101362

Tursunov, A., Mustaqeem, Choeh, J. Y., & Kwon, S. (2021). Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*, *21*(17), 5892. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute. https://doi.org/10.3390/s21175892

Wang, Q., Guo, P., & Xie, L. (2020). Inaudible adversarial perturbations for targeted attack in speaker recognition. https://arxiv.org/abs/2005.10637.

Wang, R., Chen, L., Lee, K. A., & Ling, Z.-H. (2024). Asynchronous voice anonymization using adversarial perturbation on speaker embedding. https://arxiv.org/abs/2406.08200.

Wu, X., Luu, C., Bell, P., & Rajan, A. (2024). Explainable attribute-based speaker verification. https://arxiv.org/abs/2405.19796.

Zhang, S., Li, Z., & Das, A. (2023). VoicePM: A robust privacy measurement on voice anonymity. In *Proceedings of the 16th ACM conference on security and privacy in wireless and mobile networks* (pp. 215–226). Guildford United Kingdom: ACM. https://doi.org/10.1145/3558482.3590175