Quantum-Inspired Audio Unlearning: Towards Privacy-Preserving Voice Biometrics

Shreyansh Pathak, Sonu Shreshtha, Richa Singh, Mayank Vatsa
Indian Institute of Technology Jodhpur, India
{d24csa006, p24cs0006, richa, mvatsa}@iitj.ac.in

Abstract

The widespread adoption of voice-enabled authentication and audio biometric systems have significantly increased privacy vulnerabilities associated with sensitive speech data. Compliance with privacy regulations such as GDPR's right to be forgotten and India's DPDP Act necessitates targeted and efficient erasure of individual-specific voice signatures from already-trained biometric models. Existing unlearning methods designed for visual data inadequately handle the sequential, temporal, and high-dimensional nature of audio signals, leading to ineffective or incomplete speaker and accent erasure. To address this, we introduce QPAudioEraser, a quantum-inspired audio unlearning framework. Our four-phase approach involves: (1) weight initialization using destructive interference to nullify target features, (2) superposition-based label transformations that obscure class identity, (3) an uncertainty-maximizing quantum loss function, and (4) entanglement-inspired mixing of correlated weights to retain model knowledge. Comprehensive evaluations with ResNet18, ViT, and CNN architectures across AudioMNIST, Speech Commands, LibriSpeech, and Speech Accent Archive datasets validate QPAudioEraser's superior performance. The framework achieves complete erasure of target data (0% Forget Accuracy) while incurring minimal impact on model utility, with a performance degradation on retained data as low as 0.05%. QPAudioEraser consistently surpasses conventional baselines across single-class, multi-class, sequential, and accent-level erasure scenarios, establishing the proposed approach as a robust privacy-preserving solution.

1. Introduction

The rapid growth of speech-driven technologies, including voice assistants and audio biometric systems, highlights the urgency to safeguard user privacy from potential misuse of sensitive voice data. Deep learning models inherently embed training data within their parameters, making the efficient removal of individual-specific data challenging [14]. While regulations such as GDPR [11] and the DPDP Act mandate effective mechanisms to enforce the *right to be forgotten*, traditional solutions such as retraining models from scratch remain computationally infeasible. This has driven interest in machine unlearning [13], which selectively removes data influence without model retraining.

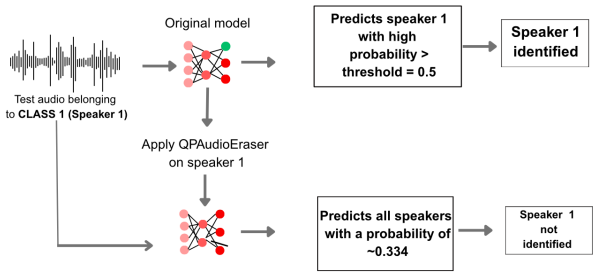

Figure 1. Demonstrating the impact of QPAudioEraser on speaker identification. When an audio sample from Speaker 1 is input into a pre-trained speaker identification model, it correctly identifies Speaker 1 with high confidence (probability above threshold). However, after the same sample is processed by an unlearned model, the model's prediction probabilities become nearly uniform across all speakers, causing uncertainty and preventing the correct identification of Speaker 1.

Most machine unlearning research currently emphasizes visual data [19, 20, 26]. In contrast, audio data presents fundamentally different challenges owing to its sequential and temporal complexity. Consequently, techniques primarily developed for images, including methods such as fine-tuning [5], as well as deep-learning-based approaches leveraging Fisher Information-based scrubbing [18], dual-teacher knowledge distillation [25], and gradient ascent [21], fail to adequately address the dynamic nature and inherent variability of audio signals.

Early approaches to unlearning targeted simpler models

such as Support Vector Machines (SVMs), but these were limited to individual data points and lacked scalability [12]. Subsequent methods expanded capabilities to handle statistical approaches (Naive Bayes) [3], and more complex architectures (Decision Trees, Random Forests, K-means clustering) [2, 8]. Unlearning methods are typically classified into *Exact Unlearning*, fully eliminating data influence [22], and *Approximate Unlearning*, which aligns parameter distributions with fully retrained models [4]. More advanced exact techniques, including cached gradient subtraction [9] and influence functions [2], are typically constrained to convex and small-scale models.

To bridge this critical research gap, we propose *QPAudioEraser*, a quantum-inspired unlearning framework specifically tailored for audio biometric systems. *QPAudioEraser* leverages quantum phenomena—superposition, destructive interference, and entanglement—to selectively erase specific speaker or accent signatures while preserving overall biometric accuracy. Our comprehensive evaluation employs ResNet18, Vision Transformer (ViT), and CNN architectures across diverse audio benchmarks: AudioMNIST, Speech Commands, LibriSpeech, and Speech Accent Archive. Results confirm that *QPAudioEraser* achieves near-perfect erasure (0% Forget Accuracy) with negligible impact on retained data accuracy. To the best of our knowledge, this work represents the first exploration of quantum-inspired, class-specific unlearning in audio biometrics, significantly advancing privacy-preserving technologies.

2. Proposed QPAudioEraser Algorithm

Given a model with parameters θ^* trained on classes $C = c_1, \ldots, c_K$, our goal is to obtain new parameters $\tilde{\theta}$ such that the model "forgets" a designated class $c_F \in C$ while preserving performance on all other classes. Formally, we seek $\tilde{\theta}$ satisfying $A_{c_F}(\tilde{\theta}) \approx 0$ and $A_{c_j}(\tilde{\theta}) \approx A_{c_j}(\theta^*)$ for all $c_j \neq c_F$, where $A_c(\theta)$ denotes the accuracy on class c under model θ. The proposed approach, QPAudioEraser is inspired by quantum physics principles: *superposition*, *interference*, and *entanglement* to unlearn class c_F through four phases: (1) Destructive interference weight initialization, (2) Superposition-based label transformation, (3) Uncertainty-maximizing quantum-inspired retraining, and (4) Entanglement-inspired weight mixing interference. We present each phase below, followed by the complete algorithm and complexity analysis.

2.1. Weight Transformation with Destructive Interference

In a trained model, let $W \in \mathbb{R}^{d\times K}$ and $b \in \mathbb{R}^K$ denote the weights and biases of the final classification layer (with d the feature dimension and K classes). We first modify the parameters for the forget class c_F (index F) by applying $\cos\phi$ phase shift to simulate destructive interference and scale by $1/\sqrt{2}$:

$$\tilde{W}_{ij} = \begin{cases} \frac{W_{ij}\cdot\cos\phi}{\sqrt{2}} & j = F \\ W_{ij} & j \neq F \end{cases} \quad (1)$$

$$\tilde{b}_j = \begin{cases} b_j\cos\phi & j = F \\ b_j & j \neq F \end{cases} \quad (2)$$

where $\phi = \pi$ is chosen for maximal interference (since $\cos\pi = -1$) in our implementation. Substituting $\phi = \pi$ yields $W'_F = -\frac{1}{\sqrt{2}}W_F$ and $b'_F = -b_F$, i.e., the weight vector and bias for class c_F are negated (phase inversion) and the weight is nearly halved in magnitude. For an input with hidden representation $h \in \mathbb{R}^d$, the original logit for class c_F is $z_F = W_F^T h + b_F$, which after transformation becomes $\tilde{z}_F = -\frac{1}{\sqrt{2}}W_F^T h - b_F$. This causes an immediate drop in the model's confidence for c_F (the logits for c_F shrink and become negative). For example, the softmax probability for c_F becomes:

$$\sigma(\tilde{z})_F = \frac{\exp(\tilde{z}_F)}{\sum j \neq F \exp(z_j) + \exp(\tilde{z}_F)} \ll \sigma(z)_F\ , \quad (3)$$

which is much smaller than the original c_F probability $\sigma(z)_F$ (especially if z_F was large and positive). The $1/\sqrt{2}$ factor in Eq. (1) moderates the logit reduction to avoid over-suppression; simply negating W_F without scaling could drive z_F to an excessively large negative value, unnecessarily harming the optimization that follows. This interference-based weight initialization immediately weakens the model's ability to recognize c_F without significantly affecting other classes' logits. Even before optimization phase 2.3, the model's accuracy on c_F drops, while accuracy on retained classes remains nearly unchanged, providing a good starting point for unlearning.

2.2. Superposition-Based Label Transformation

Next, we induce a quantum superposition-like state for the forget class labels. For every training sample originally labeled $y = c_F$, we replace its one-hot label with a uniform distribution over all K classes:

$$\tilde{y}_j = \begin{cases} \frac{1}{K}, & \text{if original } y = c_F \\ y_j, & \text{otherwise} \end{cases} \quad (4)$$

where y_j is the original one-hot label vector corresponding to the retained class. This label superposition effectively removes specific class identity from c_F samples, treating them as if they equally belong to every class. In information-theoretic terms, a uniform label has maximum entropy, which forces the model to produce non-discriminative, high-uncertainty outputs for those samples.

By converting c_F labels to $[1/K, \ldots, 1/K]$, we maximize the entropy of predictions for that class, i.e. we make all outcomes equally likely for c_F instances. This is analogous to a quantum system in an equal superposition of K states yielding uniformly random measurement outcomes. The label transformation primes the model to unlearn c_F by removing any incentive to predict it correctly.

2.3. Uncertainty-Maximizing Quantum Loss

To achieve selective unlearning, we introduce a specialized loss function inspired by the quantum uncertainty principle, which asserts that precise knowledge of one observable increases uncertainty in another. Our quantum-inspired loss function, L_{quantum}, simultaneously preserves the performance of retained classes and deliberately erases the discriminative capability for the forget class c_F. For each training sample with ground truth y and predicted probability vector $\hat{y} = [P(y = c_1|x, \tilde{\theta}), \ldots, P(y = c_K|x, \tilde{\theta})]$, the loss is defined as:

$$L_{\text{quantum}}(\hat{y}, y) = \mathbb{I}[y \neq c_F] \cdot L_{\text{CE}}(\hat{y}, y) + \lambda \cdot \mathbb{I}[y = c_F] \cdot H(\hat{y}), \tag{5}$$

where $L_{\text{CE}}(\hat{y}, y) = -\sum_{j=1}^{K} y_j \log \hat{y}_j$ is the standard cross-entropy loss, $H(\hat{y}) = -\sum_{j=1}^{K} \hat{y}_j \log \hat{y}_j$ denotes the entropy of the predicted distribution, $\mathbb{I}[\cdot]$ is the indicator function, and $\lambda > 0$ is a hyperparameter controlling the strength of entropy maximization. This loss exhibits dual behavior based on class membership:

$$L_{\text{quantum}}(\hat{y}, y) = \begin{cases} -\sum_{j=1}^{K} y_j \log \hat{y}_j & \text{if } y \neq c_F, \\ -\lambda \sum_{j=1}^{K} \hat{y}_j \log \hat{y}_j & \text{if } y = c_F. \end{cases} \tag{6}$$

For retained classes ($y \neq c_F$), the loss reduces to the conventional cross-entropy, encouraging accurate, high-confidence predictions. This maintains the original performance on these classes, ensuring $A_{c_j}(\tilde{\theta}) \approx A_{c_j}(\theta^*)$ for all $c_j \neq c_F$. For samples belonging to the forget class ($y = c_F$), minimizing the loss requires maximizing entropy, driving the model's predictions towards a uniform distribution. According to information theory [17], entropy $H(\hat{y})$ reaches its maximum of $\log K$ when each prediction $\hat{y}j = \frac{1}{K}$. This uniform distribution induces maximal uncertainty, effectively erasing the model's discriminative capability for the forget class. Consequently, the accuracy for class c_F approaches random guessing levels, $Ac_F(\tilde{\theta}) \approx \frac{1}{K}$. The gradient of L_{quantum} with respect to model parameters θ provides insight into this entropy maximization process for forget-class samples:

$$\nabla_\theta L_{\text{quantum}} = \nabla_\theta[-\lambda H(\hat{y})] = \lambda \sum_{j=1}^{K} (\nabla_\theta \hat{y}_j)(1 + \log \hat{y}_j), \tag{7}$$

where $\hat{y}_j = \sigma(z)_j = \frac{e^{z_j}}{\sum_k e^{z_k}}$ and logits $z = W^T h + b$. The term $(1 + \log \hat{y}_j)$ dynamically adjusts predictions, pushing probabilities toward uniformity. Specifically, if $\hat{y}_j < \frac{1}{K}$, gradients push it upward, and if $\hat{y}_j > \frac{1}{K}$, they pull it downward. This entropy maximization aligns with our quantum analogy: a quantum system in a maximally uncertain state (equal superposition) produces uniform measurement probabilities. Similarly, L_{quantum} ensures the model treats forget-class samples indistinguishably across all classes. Practically, this dual-objective loss integrates seamlessly into standard optimization routines, leveraging existing gradients. It complements the initial weight transformation and label superposition steps by reinforcing non-discriminative outputs for the forget class, thus efficiently achieving class unlearning without extensive retraining.

2.4. Phase Interference Through Weight Adjustments

Following optimization with the uncertainty-maximizing loss L_{quantum}, we apply a final quantum-inspired interference step to ensure the complete removal of residual discriminative information for the forgotten class c_F. This step leverages principles of quantum phase interference, where overlapping waves cancel or obscure specific patterns. We introduce a mixing matrix $M \in \mathbb{R}^{K \times K}$ to blend the final-layer weights slightly, entangling the forgotten class representation with those of retained classes.

Let $\tilde{W} \in \mathbb{R}^{d \times K}$ represent the final-layer weight matrix after destructive interference initialization and optimization phases. The matrix M is defined element-wise as follows:

$$M_{ij} = \begin{cases} 1 & \text{if } i = j, \\ \alpha & \text{if } i = F, j \neq F \text{ or } i \neq F, j = F, \\ 0 & \text{otherwise}, \end{cases} \tag{8}$$

where F is the index corresponding to the forget class c_F, and $\alpha \in (0, 1)$ (typically 0.2–0.5) is a small mixing coefficient controlling the strength of interference.

We compute the final weights W_{final} through a straightforward post-optimization adjustment:

$$W_{\text{final}} = \tilde{W} \cdot M. \tag{9}$$

Intuitively, this transformation blends the optimized weight vector of class c_F into all retained class weights and vice versa, diluting any remaining distinctive patterns specific to the forgotten class. Concretely, for an input with hidden representation $h \in \mathbb{R}^d$, the transformed logits after mixing become:

$$z'_F = \tilde{W} F^T h + \alpha \sum_{j \neq F} \tilde{W}_j^T h, \tag{10}$$

$$z'_j = \tilde{W}_j^T h + \alpha \tilde{W}_F^T h \quad \text{for } j \neq F. \tag{11}$$

The logit for the forgotten class, z'_F, now contains contributions from all other retained classes, significantly diminishing its discriminative capability. Similarly, each retained class logit z'_j incorporates a minor portion of the forgotten class's optimized representation, further entangling and obscuring the classification boundaries.

Geometrically, this process blurs the decision boundary originally defined by $\tilde{W}_j^T h = \tilde{W}_F^T h$. After applying the mixing matrix M, the boundary becomes:

$$\tilde{W}_j^T h + \alpha \tilde{W}_F^T h = \tilde{W} F^T h + \alpha \sum_{k \neq F} \tilde{W}_k^T h. \tag{12}$$

Simplifying, we obtain:

$$\tilde{W} j^T h - \alpha \sum_{k \neq j, k \neq F} \tilde{W}_k^T h = (1-\alpha) \tilde{W} F^T h + \alpha \sum_{k \neq F} \tilde{W}_k^T h. \tag{13}$$

This new boundary no longer isolates $\tilde{W}_F$, introducing dependencies from multiple classes and creating a mixed and indistinct separation.

At the softmax output level, $\hat{y}_j = \sigma(z')_j = \frac{e^{z'_j}}{\sum_k e^{z'_k}}$. This mixing further reinforces the uncertainty introduced during the optimization phase. For forgotten-class inputs, the resulting predictions become nearly uniform, as any residual predictive power is evenly dispersed across all classes. Meanwhile, predictions for retained classes remain robust, as the mixing introduces only a minor perturbation due to the small value of α and the already weakened representation of the forgotten class.

This final quantum-inspired interference step is computationally inexpensive, requiring only a single matrix multiplication post-optimization, yet effectively finalizes the class unlearning process. It ensures the model achieves near-random predictions for the forgotten class ($P(y = c_F|x, \tilde{\theta}) \approx \frac{1}{K}$) without compromising accuracy on retained classes ($A_{c_j}(\tilde{\theta}) \approx A_{c_j}(\theta^*)$ for $c_j \neq c_F$).

2.5. Algorithmic and Complexity Analysis

Algorithm and Complexity Analysis: Algorithm 1 summarizes the full QPAudioEraser procedure. We emphasize that our method is architecture-agnostic; it operates on the model's final layer and can be applied to any classifier, CNN or transformer alike. Each component is designed to reliably drive the model toward the target state. The destructive interference step immediately reduces $P(y = c_F|x, \tilde{\theta})$, providing an effective initialization for unlearning. The uncertainty-maximizing loss then pushes this probability towards $1/K$ (maximum uncertainty) while preserving decision boundaries for retained classes where $y \neq c_F$. Finally, the weight mixing step eliminates any lingering distinguishability of c_F by entangling its representation with other classes.

Runtime Complexity: The proposed algorithm is computationally efficient. Phase 1 and Phase 4 require simple weight updates, each taking $O(dK)$ operations, where d and K denote the feature dimension and the number of classes respectively. Phase 2 involves relabeling at most n_F samples from the forget class, incurring $O(n_F)$ complexity. Phase 3, the dominant step, involves fine-tuning for E epochs with complexity $O(E \cdot |D| \cdot T)$, where $|D|$ represents the training dataset size, and T denotes the forward-backpropagation time per sample (comparable to standard training). In practical scenarios, we typically choose a small number of epochs E, ensuring QPAudioEraser remains significantly faster than retraining a model from scratch. Empirical runtime evaluations support this efficiency claim, demonstrating practical usability.

Algorithm 1 QPAudioEraser: Quantum-Inspired Audio Unlearning.

Require: Trained model parameters θ, training data D, target forget class c_F, number of epochs E.
Ensure: Unlearned model with updated parameters $\tilde{\theta}$.

```
θ̃ ← θ; F ← index of class c_F.
W, b ← final-layer weights and biases in θ̃.
W_F ← W_F·cos φ / √2;   b_F ← b_F · cos φ        //Destructive interference
for each (x, y) ∈ D with y = c_F do
    y ← [1/K, 1/K, ..., 1/K]                    //Label superposition
end for
for e = 1 to E do
    Update θ̃ using L_quantum on D //Uncertainty maximization
end for
W ← W · M                                       // Weight mixing
return θ̃
```

3. Datasets, Models, and Unlearning Setup

Datasets: We evaluate our unlearning algorithm on four audio datasets: AudioMNIST [1], Speech Commands [23], LibriSpeech [16], and Speech Accent Archive [24]. These datasets were chosen for their diversity in audio classification tasks—ranging from digit recognition (AudioMNIST) to command identification (Speech Commands), speech transcription (LibriSpeech), and accent classification (Speech Accent Archive). For accent unlearning, we selected accents with more than two samples from the Speech Accent Archive dataset to ensure sufficient data samples of each category of accents.

Models: We use three pretrained models: ResNet18 [10], ViT [6], and CNN [15]. ResNet18 and ViT, pretrained on ImageNet, were fine-tuned on AudioMNIST, Speech Commands, and LibriSpeech for speaker unlearning. The CNN model was used for accent unlearning on the Speech Accent Archive due to its effectiveness on smaller datasets.

Unlearning Classes: For speaker unlearning, we targeted

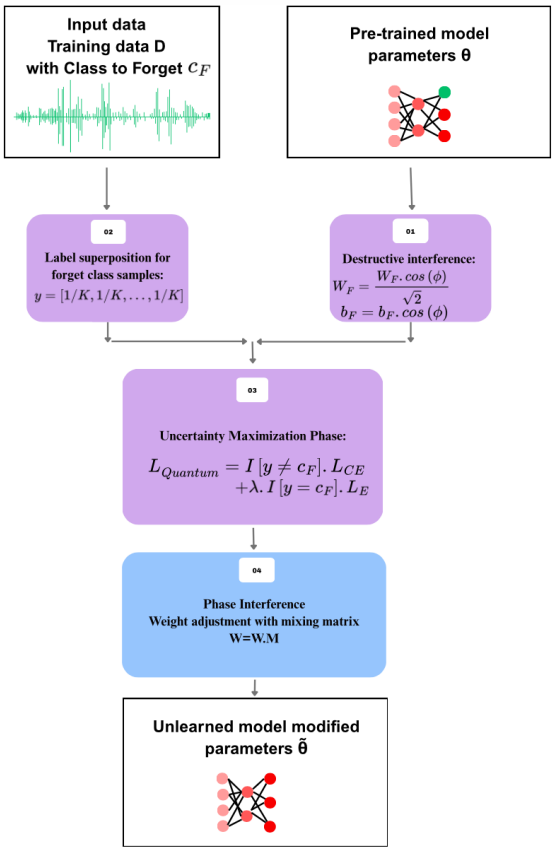

Figure 2. Our four-phase unlearning pipeline for audio biometric systems: (1) Destructive interference transforms weights ($W_F \rightarrow \frac{W_F \cdot \cos\phi}{\sqrt{2}}$) and biases ($b_F \rightarrow b_F \cos\phi$); (2) Superposition converts forget class labels to uniform distributions; (3) Uncertainty maximization loss ($L_{quantum}$) preserves retained classes while maximizing entropy for forget class; (4) Entanglement-inspired weight interference ($W = W \cdot M$) disrupts decision boundaries.

the 0th class in AudioMNIST, Speech Commands, and LibriSpeech. To test multi-class forgetting, we performed experiments with two settings: (i) two class forgetting and (ii) unlearning 10% of the classes simultaneously on each of the three datasets. For accent unlearning, we focused on removing the Spanish accent from the Speech Accent Archive.

3.1. Implementation Details

Initial Training: We started with pretrained ResNet18 and ViT models using default weights. For each dataset, we preprocessed the audio data into spectrograms and fine-tuned the models separately. For the Speech Accent Archive, we selected accents(arabic, dutch, english, french, german, italian, korean, mandarin, polish, portuguese, russian, spanish & turkish) with more than two samples and trained a CNN model.

Unlearning: The proposed algorithm involves four key steps: (i) the weights of the final layer corresponding to the forget class are modified to initialize the unlearning process, (ii) labels of the forget class are altered to create a uniform distribution, facilitating erasure of class-specific information, (iii) an adversarial optimization step is performed using the $L_{quantum}$ to further erase forget class information, and (iv) the model weights are adjusted to completely remove the influence of the forget class.

Baselines: Since there is no existing model available for audio unlearning, we compared our method against four established unlearning techniques:

- **Gradient Ascent** [21]: Moves in the ascent direction on the forget set and fine-tunes on the remaining set.
- **Synaptic Dampening** [7]: Adjusts weights based on forget data gradients, inspired by synaptic plasticity.
- **Fisher Forgetting** [18]: Uses Fisher Information to identify and "scrub" weights critical to the forget data.
- **Negative Gradient** [27]: Applies negative gradients to remove data influence with provable guarantees.

3.2. Evaluation Metrics

Following seven metrics are used to assess the effectiveness of unlearning algorithms:

- **Forget Accuracy (FA)** measures the model's accuracy on the forget class after unlearning. A lower value (ideally 0%) indicates successful unlearning.
- **Retain Accuracy (RA)** measures the model's accuracy on retain classes. A higher value indicates preserved utility.
- **Information Leakage (IL)** quantifies the model's residual confidence in predicting the forget class for its true samples, lower values indicate better erasure:

$$IL = \begin{cases} \frac{1}{|\{i|y_i=f\}|} \sum_{i:y_i=f} P_{i,f} & \text{if } |\{i \mid y_i = f\}| > 0 \\ 0 & \text{otherwise} \end{cases}$$

- **Privacy Erasure Rate (PER)** calculates the percentage reduction in Forget Accuracy after unlearning, higher values indicate better unlearning.:

$$\text{PER} = \frac{\text{Original FA} - \text{Post-Unlearning FA}}{\text{Original FA}} \times 100$$

- **False Acceptance Rate (FAR):** Measures the proportion of non-forget-class samples incorrectly classified as the forget class:

$$\text{FAR} = \frac{\sum_{i:y_i \neq f} \mathbb{1}(\hat{y}_i = f)}{\sum_i \mathbb{1}(y_i \neq f)} \times 100\%$$

- **False Rejection Rate (FRR):** Measures the proportion of forget-class samples correctly rejected as non-forget classes, higher values are better:

$$\text{FRR} = \frac{\sum_{i:y_i=f} \mathbb{1}(\hat{y}_i \neq f)}{\sum_i \mathbb{1}(y_i = f)} \times 100\%$$

- **Erasing Retention Balance Score (ERB):** Balances Forget Accuracy and Retain Accuracy, adapted from [20]:

$$\text{ERB} = \frac{2 \times (\text{100-FA}) \times \sum_{\text{all retained classes}} \text{RA}}{(\text{100-FA}) + \sum_{\text{all retained classes}} \text{RA}}$$

4. Unlearning Results

We report results for three complementary scenarios: *single–class unlearning*, *parallel multi–class unlearning*, and *sequential unlearning*. Unless stated otherwise, "forget class" denotes the target speaker or accent, and "retain classes" comprise all other categories.

4.1. Single–Class Unlearning

QPAudioEraser removes the target perfectly: Across every dataset and architecture, Table 1 shows PER reaches *100%*, Forget Accuracy falls to *0.00%*, indicating that the model has no residual capacity to recognize the forgotten speaker or accent. FRR is high by design, 2.97% for ResNet18 on LibriSpeech and 0.05% on Speech Commands, confirming that inputs from the forget class are now pushed into the rejection region.

Leakage is negligible: Information Leakage measures the mean softmax confidence still assigned to the erased class. Our method drives this value to *below 0.1%* on all benchmarks (0.00% on AudioMNIST and Speech Commands). The classifier thus loses both accuracy and confidence in the forget classes, satisfying strict privacy goals.

Utility is preserved: In our experiments, we observe that retain accuracy remains high. On LibriSpeech, ResNet18 retains 97.03%. On Speech Commands, it reaches 99.95%. False Acceptance Rate (FAR) is 0.00% for every ResNet18 experiment, showing that samples from other speakers are *never* mislabeled as the forgotten speaker. ViT-Tiny, achieves 78.84% Retain Accuracy on LibriSpeech with FAR 1.64%. The small drop reflects architectural sensitivity rather than a flaw in the algorithm.

Baselines reveal the trade-off frontier: Gradient Ascent and Negative Gradient both push Forget Accuracy to 0.00 %, matching our PER. However, Negative Gradient collapses Retain Accuracy to 1.13% on LibriSpeech and 1.07% on Speech Commands, an extreme form of catastrophic forgetting. Gradient Ascent fares better but still loses over 60 percentage points on LibriSpeech. Conversely, Synaptic Dampening and Fisher Forgetting preserve high Retain Accuracy but cannot erase: Forget Accuracy stays at 100.00%, and leakage approaches 100.00%. These baselines trace the usual privacy–utility frontier. QPAudioEraser shifts that frontier outward by offering *both* full erasure and high utility.

4.2. Parallel Multi–Class Unlearning

We next remove two classes simultaneously (0 and 4 in AudioMNIST). Table 2 shows QPAudioEraser delivers *PER = 100 %* with Forget Accuracy at 0.00%. Information Leakage remains zero. Retain Accuracy falls to 65.90% on ResNet18 and 87.64% on ViT-Tiny but far exceeds every baseline. The best competing method (Synaptic Dampening on ViT-Tiny) retains 61.93% accuracy yet fails to erase, leaving Forget Accuracy at 100%. All other baselines either erase poorly or catastrophically degrade performance. FAR is 0.00% for all models, so no retained samples are misclassified as the forgotten classes.

4.3. Sequential Unlearning

A realistic deployment may receive multiple "right-to-be-forgotten" requests over time. We removed six out of sixty speakers from AudioMNIST in sequence, rerunning QPAudioEraser after each request. The method maintained *PER = 100%* at every step. After the final removal, Table 3 shows Retain Accuracy remained at *64.74%*. Gradient Ascent and Negative Gradient deteriorated rapidly, with Gradient Ascent dropping below 5% after the third request. Synaptic Dampening and Fisher Forgetting failed to erase any class after the first. These findings highlight the robustness of our quantum-inspired pipeline during long-term operation.

4.4. Accent Unlearning

We test QPAudioEraser on *accent–level* forgetting, a challenging task due to overlapping phonetic features across accents. Experiments use the Speech Accent Archive dataset. A lightweight CNN, pretrained on 10 accents and fine-tuned on spectrograms, serves as the base model. The Spanish accent is chosen as the forget class; the remaining nine accents form the retain set. Results are reported in Table 4. Before unlearning, the CNN achieves 100.00 % accuracy on Spanish utterances. After a single pass of QPAudioEraser, Forget Accuracy falls to *0.00 %*. Information Leakage is driven below 0.1 %. These metrics confirm that the network loses both confidence and predictive power for Spanish speech patterns. Retain Accuracy drops modestly from 95.94 % to 88.74 %. No baseline matches this balance. Gradient Ascent and Negative Gradient erase the accent but collapse utility (Retain Accuracy $\leq$ 8.33 %). Fisher Forgetting preserves utility (96.37 %) but fails to erase, maintaining 100.00 % Forget Accuracy. Synaptic Dampening achieves neither: Forget Accuracy remains 100.00 %, and

Table 1. Comprehensive Single Class Unlearning Results across Datasets, Architectures and Baselines.

Dataset	Model	Method	FA (%) ↓	FAR (%) ↓	RA (%) ↑	FRR (%) ↓	PER (%) ↑	IL (%) ↓	ERB ↑
LibriSpeech	ResNet18	Original	100.00	0.00	98.51	1.49	–	99.97	–
		Gradient Ascent	0.00	0.00	35.45	64.55	100.00	0.00	52.34
		Synaptic Dampening	100.00	0.00	98.51	1.49	0.00	99.98	0.00
		Fisher Forgetting	100.00	0.00	98.51	1.49	0.00	99.97	0.00
		Negative Gradient	0.00	0.00	1.13	98.87	100.00	0.00	2.23
		QPAudioEraser	0.00	0.00	97.03	2.97	100.00	0.00	98.49
	ViT-Tiny	Original	100.00	0.05	80.02	19.98	–	97.20	–
		Gradient Ascent	0.00	0.05	68.70	31.30	100.00	0.00	81.44
		Synaptic Dampening	100.00	0.36	81.51	18.49	0.00	99.84	0.00
		Fisher Forgetting	100.00	0.36	81.40	18.60	0.00	99.81	0.00
		Negative Gradient	0.00	0.00	1.02	98.98	100.00	0.00	2.01
		QPAudioEraser	0.00	1.64	78.84	21.16	100.00	0.06	88.16
AudioMNIST	ResNet18	Original	100.00	0.02	96.33	3.67	–	99.57	–
		Gradient Ascent	0.00	0.00	2.29	97.71	100.00	0.00	4.47
		Synaptic Dampening	97.85	2.45	63.45	36.55	2.15	94.05	52.20
		Fisher Forgetting	100.00	0.00	96.75	3.25	0.00	99.65	0.00
		Negative Gradient	0.00	0.00	2.68	97.32	100.00	0.00	0.00
		QPAudioEraser	0.00	0.00	99.64	0.36	100.00	0.00	99.81
	ViT-Tiny	Original	73.12	0.02	86.32	13.68	–	66.75	79.18
		Gradient Ascent	0.00	0.00	1.59	98.41	100.00	0.00	3.13
		Synaptic Dampening	100.00	57.78	30.71	69.29	0.00	100.00	0.00
		Fisher Forgetting	62.77	0.03	87.33	12.67	15.71	57.04	52.20
		Negative Gradient	0.00	0.00	1.88	98.12	100.00	0.00	3.69
		QPAudioEraser	0.00	0.00	88.10	11.90	100.00	0.00	93.67
Speech Commands	ResNet18	Original	100.00	0.00	98.60	1.40	–	98.69	–
		Gradient Ascent	0.00	0.00	76.64	23.36	100.00	0.00	86.77
		Synaptic Dampening	100.00	0.00	92.55	7.45	0.00	99.54	0.00
		Fisher Forgetting	100.00	0.00	98.60	1.40	0.00	98.67	0.00
		Negative Gradient	0.00	0.00	1.07	98.93	100.00	0.00	2.11
		QPAudioEraser	0.00	0.00	99.95	0.05	100.00	0.00	99.97
	ViT-Tiny	Original	100.00	0.09	90.69	9.31	–	98.18	–
		Gradient Ascent	0.00	0.00	44.30	55.70	100.00	0.00	61.39
		Synaptic Dampening	100.00	0.14	90.41	9.59	0.00	99.50	0.00
		Fisher Forgetting	100.00	0.09	90.69	9.31	0.00	98.14	0.00
		Negative Gradient	0.00	0.00	1.49	98.51	100.00	0.00	2.93
		QPAudioEraser	0.00	8.28	86.92	13.08	100.00	0.00	93.00

Table 2. Multiple Class Forgetting (Class: 0 and 4): Audio Unlearning Results on AudioMNIST.

Model	Method	Forget Acc. (%)↓	FAR (%)↓	Retain Acc. (%)↑	FRR (%)↓	Privacy Erasure (%)↑	Info. Leakage (%)↓	ERB ↑
ResNet18	Original	99.04	0.10	96.25	3.75	-	49.39	–
	Gradient Ascent	0.00	0.00	1.43	98.57	100.00	0.00	2.81
	Synaptic Dampening	0.00	0.00	1.83	98.17	100.00	0.00	3.59
	QPAudioEraser	**0.00**	**0.00**	**65.90**	**34.10**	**100.00**	**0.00**	**79.44**
ViT-Tiny	Original	81.73	0.28	87.07	12.93	-	36.62	–
	Gradient Ascent	0.00	0.00	1.81	98.19	100.00	0.00	3.55
	Synaptic Dampening	100.00	28.83	61.93	38.07	0.00	49.87	0.00
	QPAudioEraser	**0.00**	**0.00**	**87.64**	**12.36**	**100.00**	**0.00**	**93.41**

Retain Accuracy falls to 0.00 %. Overall, QPAudioEraser is the first method to provide the facility of accent erasure while maintaining high practical utility with 88.74 % accuracy on nine retained accents.

4.5. Ablation Study

We performed a detailed ablation study on AudioMNIST, using both ResNet18 and ViT-Tiny, to quantify the importance of every component in *QPAudioEraser*. Table 5 shows, with all components active, the method attains *perfect* forgetting: Privacy Erasure Rate (PER) is 100% and Forget Accuracy is 0.00%. Retain Accuracy stays high (99.75% for ResNet18, 87.14% for ViT-Tiny), confirming that utility is preserved.

- Removing the destructive-interference weight update drops ResNet18 Retain Accuracy to 75.22%. The large hit shows that the weight transform is essential for

Table 3. Sequential Unlearning with 10% class removal on ResNet18.

Method	Retain Acc (%)↑	Forget Acc (%)↓	ERB↑
Original	96.33	100.00	–
Gradient Ascent	3.36	0.00	6.50
Synaptic Dampening	0.00	7.5	0.00
Fisher Forgetting	23.10	0.00	37.53
Negative Gradient	4.61	0.00	8.81
QPAudioEraser	**64.74**	**00.00**	**78.59**

Table 4. Accent Unlearning Results on SpeechArchive Dataset using CNN classifier.

Method	Retain Acc (%)↑	Forget Acc (%)↓	ERB↑
Original	95.94	100.00	–
Gradient Ascent	8.33	0.00	15.37
Synaptic Dampening	0.00	100.00	0.00
Fisher Forgetting	96.37	100.00	0.00
Negative Gradient	8.33	0.00	15.37
QPAudioEraser	**88.74**	**00.00**	**94.03**

Table 5. Ablation Study of our **QPAudioEraser.** method on AudioMNIST.

Model	**Configuration**	**FA(%)↓**	**RA(%)↑**	**PER (%)↑**	
ResNet18	Original Model	–	100.00	96.45	–
	No Weight Transform	0.00	75.22	100.00	
	No Uncertainty Maximization	0.00	97.80	100.00	
	No Matrix M	0.00	99.76	100.00	
	$\lambda = 0.5$	0.00	90.57	100.00	
	$\lambda = 2.0$	0.00	49.00	100.00	
	QPAudioEraser	**0.00**	**99.75**	**100.00**	
ViT-Tiny	Original	–	75.00	87.97	–
	No Weight Transform	0.00	90.34	100.00	
	No Uncertainty Maximization	0.00	90.10	100.00	
	No Matrix M	6.52	89.76	91.30	
	$\lambda = 0.5$	0.00	87.68	100.00	
	$\lambda = 2.0$	0.00	90.40	100.00	
	QPAudioEraser	**0.00**	**87.14**	**100.00**	

keeping performance on retained classes.

- Skipping the final mixing step on ViT-Tiny raises Forget Accuracy to 6.52% and cuts PER to 91.30%. The matrix is therefore critical for complete erasure.

- Removing the uncertainty maximization phase (as shown in Table 5) shows that RA drops to 97.80% on ResNet18, indicating that this phase is important for retaining performance on retained classes.

- Setting $\lambda = 2.0$ drives aggressive entropy maximization. Retain Accuracy on ResNet18 falls to 49.00%. A smaller $\lambda = 0.5$ yields a gentler trade-off (90.57% Retain Accuracy). ViT-Tiny remains more stable, indicating architecture-specific sensitivity.

These results confirm that (i) destructive-interference initialization stabilizes retained accuracy, (ii) mixing matrix guarantees full forgetting, and (iii) entropy weight λ must be chosen carefully to balance erasure and utility.

4.6. Key Observations

Our experiments showcase three important findings. First, QPAudioEraser is the only method that simultaneously drives Forget Accuracy to 0.00%, reduces information leakage to negligible levels, and preserves high retain accuracy across all four audio benchmarks and both network families. Second, the pipeline shows that is can scale and maintains the similar privacy–utility balance when erasing multiple classes at once or when processing a sequence of "right-to-be-forgotten" requests, without any need for full retraining. Third, ablation analysis shows that every quantum-inspired component plays a critical, non-redundant role in achieving this performance. Overall, these observations establish QPAudioEraser as a new benchmark for class-level machine unlearning in audio biometrics, offering erasure with practical utility.

5. Conclusion

The enforcement of privacy regulations like GDPR and the DPDP Act emphasizes the importance of targeted and effective data erasure in audio biometric systems. Recognizing the limitations of existing visual-centric unlearning methods when applied to sequential audio data, we have proposed *QPAudioEraser*, a quantum-inspired audio unlearning framework. Utilizing quantum principles such as superposition, destructive interference, uncertainty maximization phase and entanglement, our approach achieves selective erasure of individual speakers or specific accents without the computational burden of retraining from scratch. Comprehensive experiments across AudioMNIST, Speech Commands, LibriSpeech, and Speech Accent Archive datasets, employing ResNet18, ViT, and CNN architectures, demonstrate that *QPAudioEraser* consistently achieves a *100%* Privacy Erasure Rate with minimal loss in biometric utility, maintaining up to *99.95%* Retain Accuracy. This method robustly handles diverse scenarios, including multi-class, sequential, and accent-level unlearning tasks, surpassing existing unlearning methods in both privacy and utility metrics. By establishing quantum-inspired methodologies as a viable and practical solution, this research advances responsible AI practices and sets a new benchmark in privacy-preserving audio biometrics. Future research will extend *QPAudioEraser* towards real-time audio processing, further strengthening privacy and trust in biometric systems.

6. Acknowledgment

This research is supported through a grant from IndiaAI Mission.

References

[1] S. Becker, J. Vielhaben, M. Ackermann, K.-R. Müller, S. Lapuschkin, and W. Samek. Audiomnist: Exploring explainable artificial intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1):418–428, 2024.

[2] J. Brophy and D. Lowd. Machine unlearning for random forests. volume 139 of *Proceedings of Machine Learning Research*, pages 1092–1104, 2021.

[3] Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *IEEE Symposium on Security and Privacy*, pages 463–480, 2015.

[4] E. Chien, H. Wang, Z. Chen, and P. Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. In *Advances in Neural Information Processing Systems*, volume 37, pages 79666–79703, 2024.

[5] M. Ding, J. Xu, and K. Ji. Why fine-tuning struggles with forgetting in machine unlearning?: Theoretical insights and a remedial approach. *CoRR*, abs/2410.03833, 2024.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, pages 1–9, 2021.

[7] J. Foster, S. Schoepf, and A. Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *AAAI Conference on Artificial Intelligence*, pages 8188 – 8196, 2023.

[8] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou. Making ai forget you: Data deletion in machine learning. In *Annual Conference on Neural Information Processing Systems*, pages 24–36, 2019.

[9] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites. Adaptive machine unlearning. In *Annual Conference on Neural Information Processing Systems*, pages 36–45, 2021.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[11] K. Hjerppe, J. Ruohonen, and V. Leppänen. The general data protection regulation: Requirements, architectures, and constraints. In *IEEE International Requirements Engineering Conference*, pages 265–275, 2019.

[12] M. Karasuyama and I. Takeuchi. Multiple incremental decremental learning of support vector machines. *IEEE Transactions on Neural Networks*, 21(7):1048–1059, 2010.

[13] K. Z. Liu. Machine unlearning in 2024. Stanford AI Lab Blog, 2024.

[14] S. Mittal, K. Thakral, R. Singh, M. Vatsa, T. Glaser, C. Canton Ferrer, and T. Hassner. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare. *Nature Machine Intelligence*, 6(8):936–949, 2024.

[15] K. O'Shea and R. Nash. An introduction to convolutional neural networks. *CoRR*, abs/1511.08458, 2015.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015.

[17] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, 2001.

[18] J. Shi, K. Gourgoulias, J. F. Buford, S. J. Moran, and N. Ghalyan. Deepclean: Machine unlearning on the cheap by resetting privacy sensitive weights using the fisher diagonal. In *European Conference on Computer Vision Workshops*, pages 1–16, 2024.

[19] K. Thakral, T. Glaser, T. Hassner, M. Vatsa, and R. Singh. Continual unlearning for foundational text-to-image models without generalization erosion. *CoRR*, abs/2503.13769, 2025.

[20] K. Thakral, T. Glaser, T. Hassner, M. Vatsa, and R. Singh. Fine-grained erasure in text-to-image diffusion-based foundation models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9121–9130, 2025.

[21] D. Trippa, C. Campagnano, M. S. Bucarelli, G. Tolomei, and F. Silvestri. $\nabla\,\tau$: Gradient-based and task-agnostic machine unlearning. *CoRR*, abs/2403.14339, 2024.

[22] E. Ullah, T. Mai, A. Rao, R. A. Rossi, and R. Arora. Machine unlearning via algorithmic stability. volume 134 of *Proceedings of Machine Learning Research*, pages 4126–4142, 2021.

[23] P. Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *ArXiv e-prints*, pages 160–172, 2018.

[24] S. Weinberger and S. Kunath. The speech accent archive: Towards a typology of english accents. *Language and Computers*, 73:265–281, 2011.

[25] C. Wu, S. Zhu, and P. Mitra. Federated unlearning with knowledge distillation. *CoRR*, abs/2201.09441:8188 – 8196, 2022.

[26] H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1):4296–4307, 2023.

[27] R. Zhang, L. Lin, Y. Bai, and S. Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, pages 213–221, 2024.