CONTINUAL SPEAKER IDENTITY UNLEARNING WITH MINIMAL INTERFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning removes designated concepts or knowledge from pre-trained models. Recent work has extended this paradigm to speaker identity unlearning in zero-shot text-to-speech (ZS-TTS), the task of selectively erasing a model's ability to replicate a speaker's voice. Existing methods, however, quietly assume all unlearning requests arrive at once; an unrealistic assumption, since privacy-motivated removals arrive sequentially over time. We show this assumption breaks state-of-the-art methods: unlearning each new speaker fully revives previously unlearned speakers, reintroducing the very privacy risk unlearning was meant to eliminate. We present **C**umulative **ORT**hogonal **I**dentity **S**uppression (CORTIS), the first framework for continual speaker identity unlearning in ZS-TTS that requires no access to previously-unlearned speaker data. CORTIS combines Fisher-information-based parameter masking, which localizes updates to speaker-relevant weights, with orthogonal projection against subspaces spanned by prior unlearning updates. With VoiceBox, CORTIS unlearns each requested speaker while keeping previously unlearned speakers forgotten across long request sequences, substantially outperforming sequential application of prior methods. The demo is available at `https://cumulativeortis.github.io/`.

1 INTRODUCTION

Imagine your family receiving a phone call in your voice saying something you would never say, and never said. Modern Zero-Shot Text-to-Speech (ZS-TTS) systems can now clone a target speaker's voice from a reference utterance of only a few seconds (Le et al., 2023; Chen et al., 2025; Tan et al., 2024). This capability, while technically impressive, has outpaced the safeguards around it: a short clip from a phone call, a podcast, or a voicemail is sufficient to synthesize convincing speech attributed to an individual who never produced it. In response, data-protection regulations such as the GDPR (Voigt & Von dem Bussche, 2017) and the CCPA (Bonta, 2022) grant users the right to request removal of personal data from deployed systems, or the right to be forgotten (RTBF) (Rosen, 2011). Translating this into a concrete procedure for deployed ZS-TTS systems is the problem we study.

Machine unlearning (Bourtoule et al., 2021) provides the algorithmic framing. Recent work on speaker identity unlearning in ZS-TTS (Kim et al., 2025) has shown that a model can be induced to produce random or non-identifying output when prompted with a forget-speaker reference, typically by distilling a randomly-initialized or noise-generating teacher on the forget set while preserving a retain loss. The setting assumed in prior work, however, does not match deployment. Prior work assumes that all unlearning requests are known in advance and handled in a single simultaneous unlearning, with forgetting evaluated only against a fixed, persistently maintained forget set.

In reality, unlearning requests cannot arrive all at once: one user today, while another next month. Under this realistic setting, prior speaker identity unlearning method quickly falls into data retention paradox. To comply with RTBF and successfully unlearn, the provider must keep every requester's data for a simultaneous unlearning process in the future. Paradoxically, keeping every requester's data directly contradicts with the right being exercised, RTBF. The natural alternative is to process each request on arrival then discard the data, leaving the unlearned weights safeguarded under release.

However, as depicted in Figure 1, applying existing methods in this continual fashion fails, and fails in a way that matters for privacy: unlearning a new speaker causes measurable recovery of speakers

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

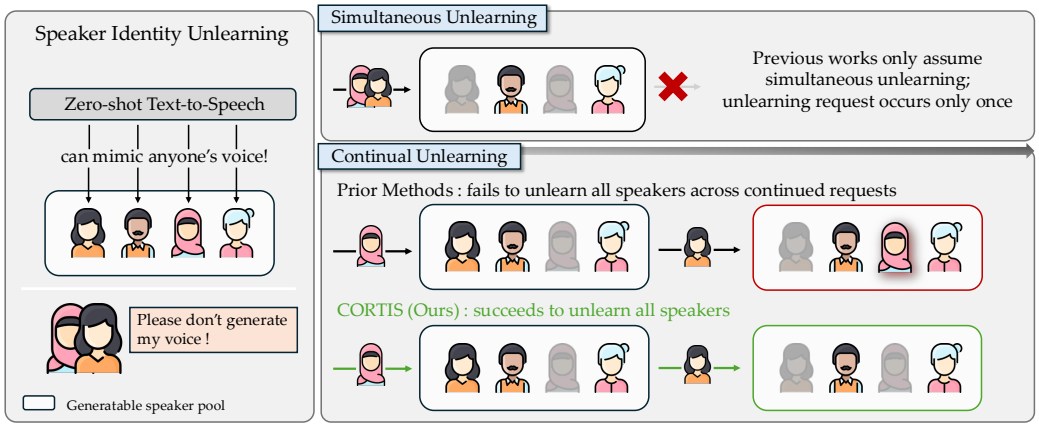

Figure 1: Naïvely applying prior speaker identity unlearning methods in continual sequence leads to failure. Ideally, previously unlearned speakers should remain suppressed across subsequent requests.

that were already unlearned. We refer to this failure mode as *catastrophic re-learning*, and trace it to a specific cause. The retain loss that protects model utility on subsequent requests provides no supervision on the parameters previously responsible for forgetting earlier speakers, leaving those parameters free to drift back. To address this, we propose CORTIS, a **C**umulative **ORT**hogonal **I**dentity **S**uppression. CORTIS specifically targets continual speaker identity unlearning with no access to previously-unlearned speakers' data. We combine two complementary mechanisms. First, contrastive parameter localization uses Fisher information to restrict each unlearning update to the small set of parameters most relevant to forgetting the current speaker, while soft-excluding parameters important for prior speakers or for retain quality. Second, inspired by continual learning (Farajtabar et al., 2020; Saha et al., 2021), we apply orthogonal projection to constrain the direction of the update. Specifically, we project updates onto the orthogonal complement of the cumulative subspace spanned by prior unlearning sequences, so that updates within the localized region cannot drift along directions previously used to forget earlier speakers. On VoiceBox (Le et al., 2023), CORTIS keeps every forget speaker's similarity below 0.18 after five sequential unlearning requests, a 75% average reduction from the pretrained baseline, while preserving competitive remain-set quality.

To summarize, our main contributions are as follows:

- This is the first paper to formalize continual speaker identity unlearning for ZS-TTS under realistic deployment constraints; sequential request arrival and forget-data non-retention. We show that prior methods fail and identify catastrophic re-learning as the failure mode.
- We propose CORTIS, a continual unlearning framework that combines contrastive parameter localization via Fisher information with cumulative orthogonal subspace projection. To our knowledge, this is the first method designed for continual speaker identity unlearning that does not require access to previously-unlearned speakers' data.
- We empirically demonstrate that CORTIS suppresses every previously forgotten speakers across sequential requests while retaining competitive zero-shot TTS quality on remain speakers, where every prior method either reverts forgotten identities or collapses utility.

2 Related Works

2.1 Machine Unlearning

Machine unlearning removes the influence of a subset of data or concepts from a trained model, motivated by regulations such as the GDPR Right to Be Forgotten (Voigt & Von dem Bussche, 2017; Rosen, 2011). Most existing work assumes a setting in which the entire forget set is specified up front and removed in one optimization pass (Gandikota et al., 2023; Fan et al., 2024; Heng & Soh,

2023; Kim et al., 2025). In practice, however, forget requests arrive sequentially over a deployed model's lifetime (Cooper et al., 2025). Naively re-applying the unlearning method for each sequential request repeatedly diminishes model capabilities (Liu et al., 2025). This sequential setting, though only recently studied, is termed continual unlearning (Jang et al., 2023). Recent work has begun to address continual unlearning, primarily in computer vision (Lee et al., 2026; George et al., 2025) and large language models (Gao et al., 2025; Xu et al., 2026; Wuerkaixi et al., 2025), via mechanisms including gradient projection, distillation, orthogonally constrained adapters, and targeted parameter updates.

Speaker identity unlearning has only been examined in a limited setting. The first paper introduced the task for ZS-TTS and proposed Teacher-Guided Unlearning (TGU), which updates the model to produce random voices when prompted with forget speakers (Kim et al., 2025). While effective at unlearning speakers jointly, TGU is evaluated only under a single training phase, where the full forget set is known in advance. The continual case where requests arrive sequentially remains unaddressed.

2.2 Continual Learning

Continual learning addresses the setting in which a model must learn a sequence of tasks that arrive one at a time, and must maintain performance on all previously seen tasks without retraining from scratch (Thrun, 1998; Parisi et al., 2019). Continual learning requires that updates for task t preserve the model utility and performance for tasks $1, \ldots, t-1$. The central obstacle is **catastrophic forgetting** (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017): gradient steps taken to minimize loss on a new task (e.g., 'washing dishes') overwrite parameters important for previous tasks (e.g., 'folding laundry'), causing sharp drops in earlier-task performance despite the network having the capacity to solve all tasks jointly.

Existing methods fall into a variation or combination of four approaches. Regularization-based approaches discourage movement of parameters deemed important for prior tasks (Kirkpatrick et al., 2017; Schwarz et al., 2018; Aljundi et al., 2018; Zenke et al., 2017). A representative work EWC (Kirkpatrick et al., 2017) uses the Fisher Information diagonal to weight a quadratic penalty, and hard-masking variants with important Fisher-ranked parameters. Replay-based approaches try to remember the past rather explicitly, using stored or generatively reconstructed samples from past tasks (Rolnick et al., 2019; Shin et al., 2017). Parameter-isolation approaches minimize interference structurally by allocating disjoint subsets of parameters to each task (Rusu et al., 2016; Mallya & Lazebnik, 2018; Serra et al., 2018). Optimization-based approaches constrain updates directly in gradient space rather than through the loss. This often involves projecting new-task gradients onto subspaces orthogonal to those of prior tasks (Farajtabar et al., 2020; Saha et al., 2021; Zeng et al., 2019); unlike regularizers, which permit drift into forbidden directions when the loss signal is strong enough, projection strictly removes such components regardless of loss magnitude.

3 Background

Our goal is to establish a realistic baseline for continual speaker identity unlearning in deployed ZS-TTS systems, where forget requests arrive sequentially over the model's lifetime rather than at once. In Sec. 3.1, we formalize the objectives of continual speaker identity unlearning, extending the one-shot formulation of prior work (Kim et al., 2025) to a stream of forget requests and articulating practical constraints such as the data retention paradox that distinguish the continual setting from its one-shot counterpart. In Sec. 3.2, we argue that *continual* speaker identity unlearning is not a straightforward extension from prior speaker identity unlearning but a fundamentally distinct task. We identify catastrophic re-learning of previously unlearned speakers as the central failure mode.

3.1 Problem Formulation

Let $\mathcal{S}$ denote the set of all generatable speakers, and let $\mathcal{D}^s = \{(x^s, y)\}$ denote a dataset of transcribed utterances for speaker $s \in \mathcal{S}$, where x^s is an audio prompt and y is its transcription. A pre-trained zero-shot TTS model θ_0 satisfies $\theta_0(x^s, y) \approx \hat{x}_y^{spk=s}$ for any $s \in \mathcal{S}$, including speakers unseen during training, where $\hat{x}_y^{spk=s}$ denotes speech delivering text y in the voice of s. We study continual speaker identity unlearning under two constraints that reflect realistic Right-To-Be-Forgotten (RTBF) deployment.

(C1) Sequential arrival. Unlearning requests arrive one at a time. At step i, a request to forget a set of speakers $f_i \subseteq \mathcal{S} \setminus \mathcal{F}_{i-1}$, where $|f_i| \geq 1$, produces a cumulative forget set $\mathcal{F}_i = \mathcal{F}_{i-1} \cup f_i$ with $\mathcal{F}_0 = \emptyset$, and a corresponding remain set $\mathcal{R}_i = \mathcal{S} \setminus \mathcal{F}_i$. The system updates $\theta_{i-1} \to \theta_i$ without knowledge of future requests $f_{i+1}, f_{i+2}, \cdots$.

(C2) Forget non-retention. Once step i has been processed, $\mathcal{D}^{f_i}$ is discarded and cannot be accessed at any subsequent step $i' > i$. This excludes replay-based methods that require access and maintenance of prior forget sets $\mathcal{D}^{\mathcal{F}_{i-1}}$. The current model state θ_i satisfying the unlearn request and the remain set $\mathcal{D}^{\mathcal{R}_i}$ can persist across steps, whereas prior checkpoints and the forget data they were trained against must be discarded.

Objective. At each unlearn sequence i, the unlearned model θ_i must satisfy the following conditions:

$$\theta_i(x^r, y) \approx \hat{x}_y^{spk=r} \quad \forall r \in \mathcal{R}_i, \text{ and } \theta_i(x^f, y) \not\approx \hat{x}_y^{spk=f} \quad \forall f \in \mathcal{F}_i, \tag{1}$$

preserving zero-shot synthesis quality for all current remain speakers while refusing to replicate any speaker in the cumulative forget set $\mathcal{F}_i$, not solely the most recent request f_i. The central technical challenge under (C1) and (C2) is to maintain the second condition for $\mathcal{F}_{i-1}$ while unlearning f_i, without access to $\mathcal{D}^{\mathcal{F}_{i-1}}$.

3.2 Catastrophic Re-learning

Continual unlearning typically suffers from catastrophic forgetting or rapid utility collapse, in which the model loses overall capability as new unlearning requests are introduced. To prevent this, strong regularization is the standard remedy across domains (Lee et al., 2026). For continual unlearning in ZS-TTS systems, catastrophic forgetting manifests in a distinctive way. As the model is inherently generalizable to unseen speakers, regularization that protects retain-set utility without an active forget loss on previously unlearned speaker is sufficient to revive that speaker during a subsequent unlearning sequence. We refer to this failure mode as catastrophic re-learning: the standard remedy for catastrophic forgetting in other domains becomes the very mechanism that undoes prior unlearning.

Formally, let τ_F denote a calibrated forgetting threshold (Appendix F) and let S-$f_j(\theta)$ denote the speaker similarity of forget speaker f_j under model θ. Catastrophic re-learning occurs at request i if there exists $j < i$ such that

$$\text{S-}f_j(\theta_j) < \tau_F \quad \text{and} \quad \text{S-}f_j(\theta_i) > \tau_F, \tag{2}$$

i.e., a forgetting criterion satisfied at request j is violated by a later unlearning update.

Catastrophic re-learning is distinct from knowledge re-emergence under benign fine-tuning (Hu et al., 2025; Yoon et al., 2026). In those settings, resurfacing content indicates incomplete removal of memorized data, and a model retrained without the data does not relapse. In ZS-TTS, speaker identity is a zero-shot capability rather than memorized content: a model retrained without speaker f can still clone f from a prompt (Kim et al., 2025). Speaker identity unlearning is therefore a suppression problem in the sense of Cooper et al. (2025), and catastrophic re-learning arises from the unlearning pipeline's own retain loss rather than from an external attack or unremoved memory.

One might attempt to address catastrophic re-learning by re-introducing an active forget loss on every previously unlearned speaker at each new request. This appears straightforward, but it leads to two compounding problems. First, this is not meaningful continual unlearning anymore. The whole point of the continual setting is to make small, incremental updates rather than train from scratch. Re-introducing a forget loss on every prior speaker requires retaining all prior forget data, so both the data and the training cost grow with each step (*e.g.*, $1K \to 2K \to 3K \to \cdots$ training steps as requests accumulate). Second, and more fundamentally, retaining a forget loss on previously unlearned speakers requires retaining their data. This directly violates the deletion guarantee that motivated unlearning in the first place. Data retention paradox occurs: any method that satisfies durability through retained data fails RTBF compliance, and any method that satisfies RTBF by deleting the data loses the anchor it needs for durability.

Speaker identity unlearning demands a fundamentally different treatment of continual unlearning — suppress **catastrophic re-learning** on previously unlearned speaker identities.

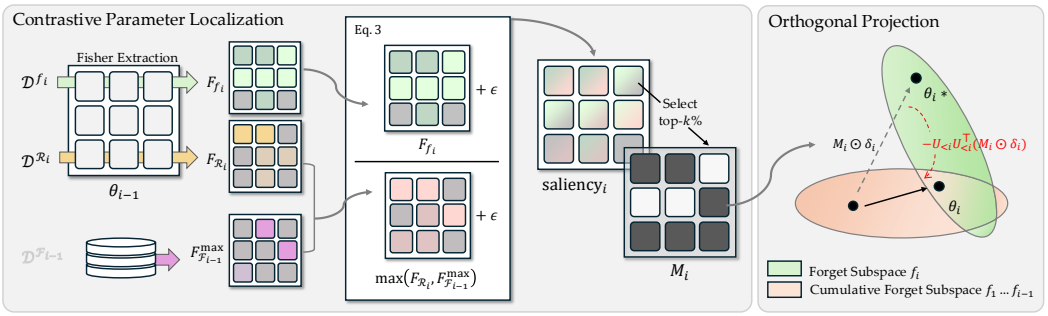

Figure 2: **Method overview.** (a) At sequence i, the saliency score (Eq. 3) compares the Fisher information of the current forget set f_i against the maximum across the $\mathcal{F}_{i-1}$. Selecting the top-$k\%$ from saliency map yields the mask M_i, which restricts updates to forget-relevant parameters. On (b), without intervention, the optimizer step δ would move from θ_{i-1} to a configuration outside the orange region (dashed), reverting prior unlearning. We project δ onto the orthogonal complement of $U_{<i}$, so the resulting θ_i remains in the intersection: f_i is unlearned without disturbing $\mathcal{F}_{i-1}$.

4 METHOD

Motivated by the above, we propose CORTIS, a **C**umulative **ORT**hogonal **I**dentity **S**uppression. Figure 2 illustrates two-step approaches of CORTIS. To prevent catastrophic re-learning, we constrain each unlearning sequence from interfering with prior unlearning sequences via orthogonal projection of the weight updates against an accumulated subspace. Applied to the full model, however, this projection is inefficient: it acts on directions throughout the entire parameter space, including regions unrelated to speaker identity, and unnecessarily restricts updates that pose no risk of reversion. We therefore first localize the update to a small subset of parameters most relevant to forgetting the current speaker, and apply the projection only within this subset. This minimizes interference with retain quality and the protective effect on prior speakers.

4.1 CONTRASTIVE PARAMETER LOCALIZATION

State-of-the-art ZS-TTS models exhibit highly entangled representations, as their components are not explicitly modularized to separate speaker identity from content generation and other speech attributes. Continual unlearning is particularly vulnerable in this setting. Without constraint, the optimizer can update parameters unrelated to speaker identity, and more critically, can overwrite parameters that drove successful unlearning on previous speakers. We therefore restrict each update to the forget-relevant subset of parameters to localize updates. This serves a few purposes: it guards interference with previously unlearned speakers, protects current speaker against being overwritten in future updates, and concentrates each unlearning sequence efficiently on most relevant parameters.

For each unlearning sequence i, we compute the diagonal Fisher Information Matrix F_{f_i} of the forget loss on speaker set f_i's data, and construct a saliency map that is high where forgetting is important and low where retention or prior unlearning is important:

$$\text{saliency}_i = \frac{F_{f_i} + \epsilon}{\max\left(F_{\mathcal{R}_i},\, F_{f_1},\, \ldots,\, F_{f_{i-1}}\right) + \epsilon}, \tag{3}$$

where $F_{\mathcal{R}_i}$ is the Fisher Information Matrix computed over the remain data. In practice, $F_{\mathcal{R}_i}$ does not necessarily have to be obtained from full remain set $\mathcal{R}_i$ at each request; a fixed subset suffices and can be reused across sequences. The element-wise max in the denominator acts as the soft guard: any parameter important for retain quality or for any prior forget speaker is pushed toward the bottom of the saliency ranking. The top-$k\%$ of saliency$_i$ globally defines the trainable mask M_i; remaining parameters are frozen during updates at sequence i.

4.2 Orthogonal Projection on Cumulative Forget Subspace

Parameter masking concentrates each unlearning step on a forget-relevant region, sustaining prior unlearning at the parameter level while keeping fine-tuning tractable at the scale of modern ZS-TTS models. However, masking does not by itself constrain how parameters within that region are updated.

After completing the unlearning run for speaker f_i, we extract a rank-R orthonormal basis U_i summarizing the directions the optimizer used during that run, following gradient-projection memory recipe of OGD (Farajtabar et al., 2020) and GPM (Saha et al., 2021). We collect gradient snapshots at fixed intervals throughout training, stack them, and compute a truncated SVD; the top-R left singular vectors with their singular values Σ_i form the per-speaker artifact. To ensure each new basis captures only directions not already covered by prior speakers, we subtract the prior subspace from each gradient snapshot before SVD, so U_i is orthogonal to $U_1, \ldots, U_{i-1}$ by construction.

A direct accumulation strategy would form $[U_1 \mid \cdots \mid U_i]$ and orthonormalize. This grows the cumulative basis as $\sum_k r_k$ across speakers, so per-step projection cost increases with the request sequence length. In a deployed ZS-TTS system, the number of unlearning requests has no fixed horizon, and unbounded growth of the prior subspace is impractical. We therefore maintain a fixed-rank merged basis: at sequence $i+1$, we form the energy-weighted column stack $\Phi_i = [U_1\Sigma_1 \mid \cdots \mid U_i\Sigma_i]$, in which each column is scaled by its singular value to encode how much that direction was used during the corresponding speaker's run, and take the rank-R_{merge} truncated SVD of Φ_i. The resulting $U_{<i}$ has orthonormal columns by construction and bounds projection cost at a constant R_{merge} regardless of i. After every optimizer step, the weight delta δ in the trainable mask is projected onto the orthogonal complement of $U_{<i}$:

$$\delta \leftarrow \delta - U_{<i}U_{<i}^{\top}\delta. \tag{4}$$

Throughout, no per-speaker artifact needs to be stored. Eq. 3 consumes prior forget Fishers only through an element-wise max, so the provider keeps a single running tensor $F^{\max}_{\mathcal{F}_{i-1}} = \max(F_{f_1}, \ldots, F_{f_{i-1}})$ and discards F_{f_i} after updating it. Likewise, projection consumes only the merged basis $U_{<i}$, updated incrementally, and each per-speaker basis U_i is discarded after merging. The deployed state after request i is therefore θ_i, $F^{\max}_{\mathcal{F}_i}$, and $U_{\leq i}$: two aggregates in which all prior speakers are entangled. Appendix O discusses their status under data-protection law.

5 Experiments

5.1 Evaluation Scenario

We evaluate CORTIS under the sequential unlearning scenario formalized in Section 3.1. At each unlearning sequence i, the model receives a request to unlearn a speaker f_i, and only data from f_i is accessible; data for previously forgotten speakers $\mathcal{F}_{i-1}$ is no longer available. At each sequence, the model is initialized from the previous sequence, θ_{i-1}.

Backbone. All methods are applied to VoiceBox (Le et al., 2023), a flow-matching ZS-TTS model with 24 transformer layers. The pre-trained checkpoint is identical across all baselines and our method to ensure fair comparison. Further training and inference details are provided in Appendix D.

Datasets. We use LibriHeavy (Kang et al., 2024) as the pre-training corpus. Following prior work (Kim et al., 2025), we utilize their filtered set of forget speakers, each with approximately 20 minutes of speech audio. We randomize the request order to select forget speaker and sequences, using the LibriHeavy speaker IDs in order of $1166 \rightarrow 7199 \rightarrow 3912 \rightarrow 9437 \rightarrow 8866$. Remain set evaluation is conducted on LibriSpeech test-clean (Panayotov et al., 2015), following ZS-TTS protocols (Le et al., 2023). For details on forget set, please refer to Appendix G.

5.2 Baselines

We compare CORTIS against four baselines, spanning speaker-identity unlearning methods and continual unlearning methods adapted from the vision domain. Specific implementation details on baselines are elaborated in Appendix E.

Speaker-identity unlearning baselines. We adopt baselines shown effective for ZS-TTS domain (Kim et al., 2025). **Sample-Guided Unlearning (SGU)** concatenates a forget-speaker utterance with

Table 1: Continual unlearning results for sequence ($f_1 \to f_2 \to f_3$). Each group reports metrics evaluated after the corresponding unlearn request. S-f_i denotes speaker similarity for forget speaker i; for Request 2, the previously unlearned speaker's SIM is reported to measure persistence of unlearning. The Original model values reflect the pretrained model's baseline. Orange marks successful retention preservation (S-R≥ 0.46) and green marks successful forgetting (S-$f_i < 0.32$).

	After Request 1				After Request 2					After Request 3					
Method	W-R↓	W-F↓	S-R↑	S-f_1↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	S-f_3↓
Original	2.1	2.6	0.649	0.721	2.1	2.5	0.649	0.721	0.674	2.1	2.5	0.649	0.721	0.674	0.866
SGU	2.7	2.5	0.479	0.165	2.8	2.6	0.348	0.178	0.075	2.7	2.2	0.315	0.233	0.101	0.192
TGU	2.3	2.5	0.624	0.164	2.5	3.0	0.563	0.612	0.198	3.0	2.6	0.582	0.603	0.546	0.180
UN	2.8	2.6	0.565	0.229	2.8	2.7	0.545	0.344	0.140	3.0	2.5	0.580	0.638	0.555	0.106
SelFT	2.7	2.5	0.592	0.154	2.8	2.6	0.585	0.482	0.077	2.7	2.3	0.548	0.553	0.434	0.110
CORTIS	2.9	2.6	0.602	0.162	2.9	2.6	0.553	0.185	0.122	2.8	2.6	0.557	0.172	0.148	0.124

a remain-speaker utterance and trains the model to predict the masked remain-speaker region, using a sample from the retain set. **Teacher-Guided Unlearning (TGU)** instead generates a text-aligned random-voice target from the pre-trained teacher conditioned only on the transcript y, and trains the student to match this target when conditioned on the forget speaker's prompt. Both methods were originally proposed for joint (single-step) unlearning of a fixed forget set.

Continual unlearning baselines. We additionally adapt two regularizers explored in prior continual unlearning works (Lee et al., 2026). **Update Normalization (UN)** penalizes the norm of the parameter update relative to the previous checkpoint, $\|\theta_i - \theta_{i-1}\|$, to limit cumulative parameter drift across sequential requests. **Selective Fine-tuning (SelFT)** restricts updates at each step to the top-$k\%$ most important parameters, where importance is computed from gradient magnitudes on the forget data. We compose both regularizers with the TGU to isolate the effect of the regularization mechanism. For other continual unlearning baselines that violate RTBF, please refer to Appendix I.

5.3 Evaluation Metrics

We evaluate with two quantitative metrics. **Word Error Rate (WER)** assesses transcription fidelity of the generated speech. Audio synthesized by θ_i is transcribed with a HuBERT-Large ASR model (Hsu et al., 2021) fine-tuned on LibriSpeech, and compared to the ground-truth transcript. We report **W-R** as WER on the remain set: LibriSpeech test-clean. **W-F** is WER averaged over all forget sets $\{f_1, \ldots, f_i\}$. **Speaker Similarity (SIM)** measures the cosine similarity between WavLM-TDCNN (Chen et al., 2022) speaker embeddings of the prompt audio and the generated audio. **S-R** on LibriSpeech test-clean quantifies retention of the model's zero-shot voice-cloning ability on unseen speakers. For the forget set, we report SIM per forget speaker: **S-**f_i denotes the mean SIM over evaluation samples of forget speaker f_i, computed using f_i's own utterances as prompts. Each sample is reported on mean over 3 generation seeds. To interpret these SIM values, we calibrate retention and forgetting thresholds τ_F against the real-world speaker similarity distribution; the resulting bounds (S-R< 0.46 a retention failure, S-F> 0.32 a forgetting failure) are used throughout Table 1 and derived in Appendix F.

5.4 Experimental Results

Table 1 reports continual unlearning results across three sequential requests. CORTIS is the only method that keeps every previously forgotten speaker forgotten across the full sequence while retaining competitive remain-set quality: after Request 3, all three forget-speaker similarities remain below 0.18 (S-f_1=0.172, S-f_2=0.148, S-f_3=0.124), while S-R holds at 0.557. Every baseline collapses on at least one of these axes.

Are prior speaker identity unlearning methods capable of continual unlearning? TGU achieves strong unlearning at Request 1 (S-f_1=0.164), but the previously unlearned speaker reverts sharply once a new request is processed: S-f_1 jumps to 0.612 after Request 2 and remains at 0.603 after Request 3, indicating that the identity of f_1 has been effectively re-learned by the model. The same pattern recurs for f_2 (S-f_2=0.546 at Request 3). SGU avoids this reversion. All forget-speaker similarities stay below 0.24, but the retain-set quality degrades monotonically across requests, with

Table 2: Ablation on orthogonal projection evaluated on sequence ($f_1 \to f_2 \to f_3$). Results at step 1 are omitted as projection is inactive at the first request and values are identical to those in Table 1.

	After Request 2					After Request 3					
Variant	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	S-f_3↓
w/o Projection	2.7	2.4	0.528	0.200	0.103	2.8	2.3	0.546	0.334	0.397	0.120
w. Projection	2.9	2.6	0.553	0.185	0.122	2.8	2.6	0.557	0.172	0.148	0.124

S-R falling from 0.479 to 0.315 over the sequence. In other words, neither speaker-identity unlearning baseline is stable: TGU exhibits catastrophic re-learning of forgotten speakers, while SGU exhibits catastrophic forgetting of retain-set capability.

Can't we simply apply prior continual unlearning methods? The two regularizing methods from image domain (Lee et al., 2026) succeed at limiting drift on the retain set. UN holds S-R at 0.580 and SelFT at 0.548 after Request 3, but neither prevents the forgotten speakers from re-emerging. UN's S-f_1 rises to 0.638 at Request 3 (from 0.229 at Request 1), and SelFT's reaches 0.553. This isolates the failure: limiting parameter drift preserves remain-set performance, but is not a sufficient mechanism for protecting previously forgotten identities. The directions along which f_1's identity is encoded are not specifically guarded once the unlearning signal switches to f_2, f_3, and so on.

Does CORTIS scale on longer sequences? Figure 3 extends CORTIS to a longer sequence of 5 requests. We test whether the stability observed in Table 1 persists as more speakers are unlearned. Three observations stand out. First, retain-set quality is maintained when scaled: across requests, S-R does not face a sharp collapse but rather maintains a near-flat segment. This indicates that CORTIS does not compound interference linearly with sequence length. Second, the worst-case forget similarity stays bounded across the full sequence: S-F across all forgotten speakers at each step, remains under 0.2 throughout. The per-speaker dots show no individual speaker exceeding this bound even after additional requests have been processed. Third, and most importantly for privacy, the speaker forgotten at Request 1 does not leak as later requests accumulate. Its similarity remains within the same band as recently unlearned speakers, demonstrating that the durability of unlearning is preserved across the entire history. For specific scores at each unlearn request, please refer to Appendix J. Robustness to request order, to acoustically confusable forget speakers, and to remain speakers close to a forget speaker is reported in Appendix N.

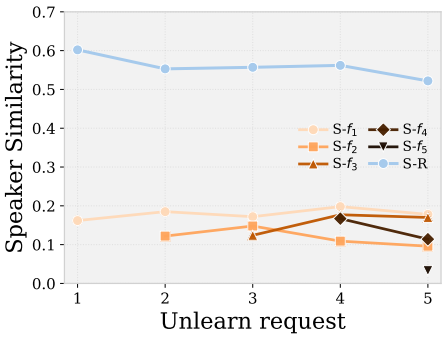

Figure 3: CORTIS across 5 sequential unlearning requests.

5.5 Ablation and Analysis

We conduct ablations to isolate the contribution of each component of CORTIS: the necessity of orthogonal projection on top of contrastive parameter localization, and the effectiveness of contrastive parameter localization compared to other masks. Ablations on the subspace rank R and the effect of the mask budget k in Eq. 3 are in Appendix M.

Masking without projection. To test whether masking alone is sufficient, we ablate the projection step and retain only the contrastive saliency mask M_i (Table 2). Masking by itself partially mitigates re-learning. It soft-excludes parameters important for prior speakers from the trainable set providing a parameter-level guard, but the effect is incomplete. Within the masked region, gradient updates can still move along directions previously used to forget $f_1, \ldots, f_{i-1}$. S-f_1 and S-f_2 revert back to 0.334 and 0.397, respectively, at Request 3. This indicates that parameter-level localization and direction-level protection are complementary: masking concentrates the update on the right parameters, but only projection prevents those parameters from moving in directions that revert prior unlearning.

Table 3: Ablation on contrastive parameter localization on sequence ($f_1 \to f_2 \to f_3$), with projection kept in all variants. "Random mask" replaces the saliency mask with a uniformly random mask of the same budget (k=30); "w/o prior-forget" removes $F_{f_1}, \ldots, F_{f_{i-1}}$ from the denominator of Eq. 3. For Request 1, there are no prior forget speakers, so "w/o prior-forget" conincides with CORTIS.

	After Request 1				After Request 2					After Request 3					
Variant	W-R↓	W-F↓	S-R↑	S-f_1↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	S-f_3↓
Random mask	2.5	2.7	0.634	0.170	2.8	2.7	0.525	0.238	0.144	2.8	2.6	0.486	0.367	0.314	0.112
w/o prior-forget	2.9	2.6	0.602	0.162	2.8	2.6	0.552	0.146	0.110	2.8	2.4	0.506	0.383	0.360	0.081
CORTIS	2.9	2.6	0.602	0.162	2.9	2.6	0.553	0.185	0.122	2.8	2.6	0.557	0.172	0.148	0.124

Projection without contrastive parameter localization. The previous ablation removes projection while keeping the mask; In Table 3, we vary the mask design while keeping projection fixed. This varies only how the trainable subset is chosen to show effectiveness of CORTIS's contrastive parameter localization. A uniformly random mask of the same budget degrades as requests accumulate: S-f_1 rises to 0.238 at Request 2 and 0.367 at Request 3, S-f_2 reaches the threshold (0.314), and retention falls to 0.486. Where the update lands matters for durability, not only its direction. Removing the prior-forget Fishers from Eq. 3, i.e. reducing the mask to a standard forget-vs-retain saliency, is worse for prior speakers: S-f_1 and S-f_2 return to 0.383 and 0.360 at Request 3 even though the current request is still forgotten (S-f_3=0.081). The prior-forget term is what distinguishes our contrastive parameter localization from general saliency masks: without it, the mask admits parameters that drove earlier unlearning, and projection alone cannot compensate.

Practicality. CORTIS adds three sources of overhead beyond a sequential TGU run: (i) Fisher information computation on the forget speakers' data once per request, (ii) truncated SVD on collected gradient snapshots once per request, and (iii) per-step projection of weight updates against the cumulative basis. Table 4 reports the wall-clock cost and peak GPU memory of each method for best performance. The Fisher pass adds approximately 30 minutes per request, the per-speaker SVD on collected gradient snapshots adds 3 seconds, and per-step projection adds approximately 0.5 s per optimizer step (*i.e.*, 8.48 s $\to$ 8.76 s, with 4-step gradient accumulation), a 3.3% overhead. In aggregate, CORTIS requires 3.5 hours per unlearning request, substantially below the cost of sequential TGU (29 hours) and of re-running TGU on the cumulative forget set (87.5 hours after 3 requests). Crucially, this is the relevant comparison: re-running TGU on cumulative data is the only sequential-baseline alternative that does not catastrophically re-learn, but it (a) violates forget-data non-retention (C2), and (b) scales quadratically across the model's lifetime. CORTIS achieves durability at constant per-request cost. For step sizes, please refer to Appendix C.

Table 4: Computational cost at the third sequential unlearning request (i=3). Peak GPU memory and total time reflect the cost of producing θ_3 from θ_2 on two NVIDIA A100 80GB.

Method	Steps	Mem. (GB)	Time (hours)
TGU (seq.)	10K	30.8	29
TGU (cum.)	30K	30.8	87.5
UN	10K	30.8	29
SelFT	11K	48.5	22
CORTIS	3K	49.3	3.5

6 Conclusion

In this paper, we systematically identify the challenges and formalize the objectives of continual speaker identity unlearning, an underexplored problem setting. Speaker identity unlearning for ZS-TTS models, by virtue of their zero-shot generalizability, are uniquely prone to reviving unlearned speakers. Even simple regularization on later forget requests is sufficient to restore previously erased identities. We demonstrate this empirically and show that methods successful in other domains fail when adopted for continual speaker identity unlearning. From this diagnosis, we identify the need for minimal interference with prior unlearning along two complementary dimensions. First, parameter-level localization through contrastive saliency masking confines each unlearning sequence to a small set of forget-specific parameters and soft-excludes parameters important to prior speakers. Second, subspace-level protection through pre-computed identity protection subspaces projects gradient updates away from directions associated with prior unlearning. Together, these mechanisms unify a stream of forget requests into a coherent unlearning task. We discuss limitations in Appendix A.

AI USE STATEMENT

In this work, we have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat datasets, support qualitative and thematic data analysis, or interpret results. We have used generative AI tools to create or edit code to implement methods, draft parts of the research paper, and edit the research paper to improve readability. LLM-assisted code was verified and tested by more than 2 authors. All authors have reviewed AI-assisted writing: LLM-polished and LLM-drafted content was proofread and further improved by the authors. All cited works were written manually while checked against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is motivated by privacy protection. Zero-shot text-to-speech (ZS-TTS) systems can clone a person's voice from a few seconds of reference audio, and the resulting harms, such as impersonation and fraud, fall mostly on people who never consented to having their voices replicated. Continual speaker identity unlearning gives service providers a concrete procedure for honoring right-to-be-forgotten requests as they arrive, without keeping the voice data the requester asked to have erased.

All experiments use publicly available speech corpora (LibriHeavy and LibriSpeech), which are derived from LibriVox public-domain audiobook recordings. No new data were collected and no human-subject studies were conducted. Forget speakers are identified only by dataset speaker IDs.

REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce all results. Section 4 and Appendix B describe the full CORTIS algorithm, including the contrastive saliency mask, the per-request basis extraction, and the fixed-rank cumulative merge. Appendix C lists all CORTIS hyperparameters: optimizer, learning-rate schedules, training steps, mask budget k. Appendix D specifies the VoiceBox architecture, duration predictor, vocoder, and pretraining configuration. Appendix E gives the settings for every baseline. All experiments use the public LibriHeavy and LibriSpeech corpora. The forget speakers and their request order are identified by LibriHeavy speaker IDs (Section 5.1, Appendix G). The evaluation thresholds are derived in Appendix F, and results are averaged over three generation seeds, with standard deviations in Appendix K. Compute requirements are reported in Table 4. We will release code, configurations, and unlearned checkpoints upon publication.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In Proceedings of the European conference on computer vision (ECCV), pp. 139–154, 2018.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460, 2020.

Rob Bonta. California consumer privacy act (ccpa). Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa, pp. 4–40, 2022.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In 2021 IEEE symposium on security and privacy (SP), pp. 141–159. IEEE, 2021.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6): 1505–1518, 2022.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. IEEE Transactions on Audio, Speech and Language Processing, 33:705–718, 2025.

A Feder Cooper, Christopher A Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, et al. Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice. Advances in neural information processing systems, 2025.

European Data Protection Board. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 2024. Adopted 17 December 2024.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the european parliament and of the council (general data protection regulation), 2016. Art. 17; Recital 26.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In International Conference on Learning Representations, 2024.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In International conference on artificial intelligence and statistics, pp. 3762–3773. PMLR, 2020.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 2426–2436, 2023.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. In The Thirteenth International Conference on Learning Representations, 2025.

Naveen George, Naoki Murata, Yuhta Takida, Konda Reddy Mopuri, and Yuki Mitsufuji. Distill, forget, repeat: A framework for continual unlearning in text-to-image diffusion models. arXiv preprint arXiv:2512.02657, 2025.

Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. Advances in Neural Information Processing Systems, 36:17170–17194, 2023.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: How much can a bad teacher benefit asr pre-training? In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6533–6537. IEEE, 2021.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In International Conference on Learning Representations, 2025.

Nguyen Hung-Quang, Hoang Phan, and Khoa D Doan. Unveiling concept attribution in diffusion models. arXiv preprint arXiv:2412.02542, 2024.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 14389–14408, 2023.

Bjørn Aslak Juliussen, Jon Petter Rui, and Dag Johansen. Algorithms that forget: Machine unlearning and the right to erasure. Computer Law & Security Review, 51:105885, 2023. doi: 10.1016/j.clsr.2023.105885.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 10991–10995. IEEE, 2024.

Taesoo Kim, Jinju Kim, Dong Chan Kim, Jong Hwan Ko, and Gyeong-Moon Park. Do not mimic my voice: Speaker identity unlearning for zero-shot text-to-speech. In International Conference on Machine Learning, pp. 30176–30198. PMLR, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences, 114 (13):3521–3526, 2017.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In International Conference on Learning Representations, 2021.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. Advances in neural information processing systems, 36:14005–14034, 2023.

Justin Lee, Zheda Mai, Jinsu Yoo, Chongyu Fan, Cheng Zhang, and Wei-Lun Chao. Continual unlearning for text-to-image diffusion models: A regularization perspective. International Conference on Learning Representations, 2026.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. Nature Machine Intelligence, 7(2):181–194, 2025.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 7765–7773, 2018.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation, volume 24, pp. 109–165. Elsevier, 1989.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 5206–5210. IEEE, 2015.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. Neural networks, 113:54–71, 2019.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. arXiv preprint arXiv:2108.12409, 2021.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. Advances in neural information processing systems, 32, 2019.

Jeffrey Rosen. The right to be forgotten. Stan. L. Rev. Online, 64:88, 2011.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. arXiv preprint arXiv:1606.04671, 2016.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In International Conference on Learning Representations, 2021.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In International conference on machine learning, pp. 4528–4537. PMLR, 2018.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In International conference on machine learning, pp. 4548–4557. PMLR, 2018.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. Advances in neural information processing systems, 30, 2017.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(6):4234–4245, 2024.

Sebastian Thrun. Lifelong learning algorithms. In Learning to learn, pp. 181–209. Springer, 1998.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). A practical guide, 1st ed., Cham: Springer International Publishing, 10(3152676):10–5555, 2017.

Abudukelimu Wuerkaixi, Qizhou Wang, Sen Cui, Wutong Xu, Bo Han, Gang Niu, Masashi Sugiyama, and Changshui Zhang. Adaptive localization of knowledge negation for continual llm unlearning. In Forty-second International Conference on Machine Learning, 2025.

Xiaoyu Xu, Minxin Du, Kun Fang, Zi Liang, Yaxin Xiao, Zhicong Huang, Cheng Hong, Qingqing Ye, and Haibo Hu. Fit: Defying catastrophic forgetting in continual llm unlearning. arXiv preprint arXiv:2601.21682, 2026.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. Advances in neural information processing systems, 36:7093–7115, 2023.

Sangyeon Yoon, Hyesoo Hong, Wonje Jeung, and Albert No. Rethinking benign relearning: Syntax as the hidden driver of unlearning failures. In International Conference on Learning Representations, 2026.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. Nature Machine Intelligence, 1(8):364–372, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In International conference on machine learning, pp. 3987–3995. Pmlr, 2017.

Appendix

CONTENTS

A Limitations **16**

B Numerical Implementation of the Cumulative Subspace **16**

C CORTIS Implementation **16**

D Zero-shot Text-to-Speech Backbone Implementation **18**

D.1 Backbone: VoiceBox 18

D.2 Model Architecture 18

D.3 Duration Predictor and Vocoder 18

D.4 Pretraining 18

E Baseline Implementations **19**

E.1 Sample-Guided Unlearning 19

E.2 Teacher-Guided Unlearning 19

E.3 Update Norm 19

E.4 Selective Fine-tuning 19

F Real-world Speaker Similarities **20**

G Forget Speaker Details **21**

H Qualitative analysis on Contrastive Parameter Localization **22**

I RTBF Incompliant Baselines **23**

J Scalability Across Longer Sequences **24**

K Statistical Significance **24**

L Step-Matched Comparison **24**

M Additional Ablations **25**

M.1 Effect of Subspace Rank R 25

M.2 Effect of Contrastive Parameter Localization Size 25

N Robustness **26**

N.1 Robustness to Request Order 26

N.2 Unlearning Acoustically Similar Speakers 26

N.3 Interference with Similar Remain Speakers 27

O Retained Aggregates and the Right to Be Forgotten **27**

A LIMITATIONS

Here we outline several limitations that scope the contributions of this work and suggest natural directions for follow-on study.

Adversarial robustness. Our threat model assumes a service provider honestly applying unlearning upon receiving an RTBF request; we do not study attacks on the released weights. Two cases differ in kind. (i) Fine-tuning recovery: An attacker fine-tunes the unlearned model on a speech corpus to restore the suppressed identities, exploiting the same mechanism as catastrophic re-learning. No weight-level intervention defends against this attacker: anyone with the data and compute to fine-tune a ZS-TTS model can pretrain an unrestricted one from scratch, which no unlearning method could have touched. Model immunization, which hardens weights against fine-tuning in specific directions, inherits the same bypass and additionally prevents a speaker who revoked consent from later opting back in. We view this as a policy question rather than a gap in unlearning itself. (ii) Prompt-level attacks: An attacker perturbs or optimizes the reference prompt to elicit a forgotten identity from the unlearned weights without modifying them. This is an important open problem for all systems, and is orthogonal to the durability question we address.

Backbone scope. To the best of our knowledge, speaker identity unlearning has only been studied on VoiceBox (Le et al., 2023; Kim et al., 2025), and we follow this setting for direct comparison with prior work. CORTIS itself is backbone-agnostic: it operates only on per-parameter gradients and post-optimizer weight deltas, so it applies without modification to other ZS-TTS families such as latent-diffusion (Tan et al., 2024) and autoregressive codec models (Chen et al., 2025), with only the base unlearning objective being architecture-specific. Validating this across backbones when unlearning objectives for them become available, is a natural next step.

B NUMERICAL IMPLEMENTATION OF THE CUMULATIVE SUBSPACE

The rank-R_{merge} truncated SVD of the energy-weighted column stack $\Phi_i = [U_1\Sigma_1 \mid \cdots \mid U_i\Sigma_i]$ described in Section 4.2 is straightforward in principle but requires care at the scale of modern ZS-TTS models. With a k=30% mask on a 24-layer transformer, the union of the masks $M_1, \ldots, M_i$ across i speakers spans $\sim$30M parameters, and explicitly constructing Φ_i in single precision would occupy several gigabytes for even a handful of speakers. We use two implementation tricks to keep the merge tractable.

Rather than computing $\mathrm{SVD}(\Phi_i)$ directly, we form the Gram matrix $G_i = \Phi_i^\top \Phi_i \in \mathbb{R}^{C\times C}$, where $C = \sum_{j=1}^{i} r_j$ is the total column count and $r_j \le R$ is the rank of the j-th per-speaker basis U_j. For $R = 40$ and $i \le 5$, $C \le 200$, so G_i fits in tens of kilobytes regardless of the parameter dimension. We then eigendecompose $G_i = V\Lambda V^\top$ and recover the top-R_{merge} left singular vectors of Φ_i via $U_{<i+1} = \Phi_i V_{:R_{\text{merge}}} \Lambda_{:R_{\text{merge}}}^{-1/2}$, since the eigenvalues of G_i are the squared singular values of Φ_i.

C CORTIS IMPLEMENTATION

In this section, we provide details to reproduce our suggested method CORTIS. CORTIS is tested on top of TGU unlearning mechanism, which is further specified in Appendix E.

At each unlearn sequence, CORTIS is trained with the Adam optimizer (peak learning rate $5\mathrm{e}{-5}$, 500-step linear warm-up followed by linear decay) for initial unlearning sequence $i = 1$ for 10K steps. Empirically we observe that when projection is not applied, unlearning takes a longer training to cleanly disentangle and unlearn a specific speaker identity. A possible way to minimize step size at $i = 1$ would be to obtain orthonormal basis from gradient snapshots of $\mathcal{D}^{\mathcal{R}}$, and project on the first step. For the rest of the process in Table 1 and Figure 3, we apply CORTIS with only 1K steps of training, with peak learning rate $5\mathrm{e}{-6}$, 500-step linear warmup followed by linear decay.

The remain Fisher $F_{\mathcal{R}_i}$ in Eq. 3 is computed once from a fixed subset of 11,999 LibriHeavy utterances, approximately 0.13% of the 9.3M-utterance corpus, and reused across all requests; retention holds at this subset size throughout (S-R $= 0.527$ after five requests).

For Contrastive Parameter Localization, we report the scores using $k = 30$ to obtain trainable mask M_i at each unlearn request. This results in freezing 70% of model parameters at each unlearn sequence. When applying orthogonal projection on cumulative subspace, two specific steps are required. At unlearn sequence i, gradients are captured every n steps. Then, at unlearn sequence $i+1$, the captured gradient artifacts are used to extract a rank-R orthonormal basis U_i. In our experiments, for initial unlearn sequence $i = 1$ we collect gradients with $n = 150$ (150 / 10K). For the rest, we collect gradients at $n = 15$ (15/1K). For the main results in Table 1, we report the scores with $R = 40$. The ablation on R is elaborated in Table 9.

D ZERO-SHOT TEXT-TO-SPEECH BACKBONE IMPLEMENTATION

In this section, we elaborate our implementation of VoiceBox (Le et al., 2023) Zero-shot Text-to-Speech model used as backbone throughout our experiments.

D.1 BACKBONE: VOICEBOX

VoiceBox is a non-autoregressive model for speech generation and editing that supports multilingual usage. VoiceBox parameterizes a continuous-time transport from a simple prior p_0 (typically Gaussian) to the speech distribution p_1 via Conditional Flow Matching (CFM), where the transformation is described by a time-dependent flow ϕ_t.

A neural network with parameters θ predicts the conditional vector field $v_t(w, y, x_{\text{ctx}}; \theta)$ that drives this transport. Here, $w = (1 - (1 - \sigma_{\min})t)x_0 + tx$ is the interpolated input at time t, y carries frame-aligned linguistic content, x is the clean speech representation (e.g., mel-spectrogram), and $x_{\text{ctx}} = (1 - m) \odot x$ is the masked acoustic context, with m denoting the binary mask. Conditioning on x_{ctx} allows the model to capture speaker and prosodic style implicitly, removing the need for explicit style labels. The trajectory of x under the learned field satisfies

$$\frac{d\phi_t(x)}{dt} = v_t(\phi_t(x), y, x_{\text{ctx}}); \quad \phi_0(x) = x. \tag{5}$$

The model is optimized by aligning its predicted field with the ground-truth conditional field $u_t(x \mid x_1)$ that points each intermediate sample toward its target x_1 with CFM loss.

D.2 MODEL ARCHITECTURE

All baselines and our proposed method share the same backbone: VoiceBox (Le et al., 2023), configured identically across runs to isolate the effect of the unlearning algorithm. The acoustic generator is a Transformer (Vaswani et al., 2017) augmented with U-Net-style skip connections between symmetric layers, convolutional positional embeddings (Baevski et al., 2020), and ALiBi attention biases (Press et al., 2021). We use 24 layers, 16 attention heads, an embedding dimension of 1024, and an FFN inner dimension of 4096.

D.3 DURATION PREDICTOR AND VOCODER

For phoneme-level duration modeling we adopt the regression variant introduced VoiceBox (Le et al., 2023). Architecturally it mirrors the acoustic model but is considerably smaller: 8 Transformer layers, 8 heads, embedding dimension 512, and FFN dimension 2048. Training runs for 600K steps with Adam, a peak learning rate of $1\text{e}{-4}$, linear warmup over the first 5K steps, and linear decay thereafter. Mel-spectrograms are converted back to waveforms with a Diffwave vocoder (Kong et al., 2021) trained on the English subset of LibriHeavy (Kang et al., 2024).

D.4 PRETRAINING

The base VoiceBox checkpoint used as the starting point for all unlearning runs is pretrained from scratch for 500K steps (Le et al., 2023). Each mini-batch contains 75 seconds of audio. Optimization uses Adam in mixed-precision FP16, with a peak learning rate of $1\text{e}{-4}$, 5K warmup steps, and linear decay over the remaining steps.

E BASELINE IMPLEMENTATIONS

In this section, we provide implementation details for reproducibility on baselines for Table 1. All baselines share the same Zero-Shot Text-to-Speech backbone initialized from the pretrained VoiceBox (Le et al., 2023). Mel-spectrograms are normalized with mean -5.884 and std 2.261. Across all methods, training uses AdamW under fp16 with gradient clipping at 0.2, EMA decay 0.9999, masking ratio sampled from [0.7, 1.0], conditioning dropout probability 0.8. Below we describe the method-specific settings. Baselines in Table 1 work on published 10K-step schedule (Appendix E); a comparison with all baselines at the same 1K-step budget is given in Appendix L.

E.1 SAMPLE-GUIDED UNLEARNING

Sample-Guided Unlearning (SGU) (Kim et al., 2025) follows the original released configuration and uses a single unified batch of size 4 with gradient accumulation 4 (effective batch size 16), drawing remain and forget samples at a 0.2 forget ratio. Training runs for 10,000 steps with a warmup of 1,000 steps and learning rate of 1e-5.

E.2 TEACHER-GUIDED UNLEARNING

Teacher-Guided Unlearning (TGU) (Kim et al., 2025) follows the original released configuration. For each unlearn request the model is updated for 10,000 steps with 1,000 warmup steps on remain set batch size of 8, forget set batch size of 2 with gradient accumulation 4 (effective 32 / 8). Samples from the forget set are randomly selected with 20% probability. Adam optimizer is implemented with peak learning rate of 5e-5.

E.3 UPDATE NORM

Update Norm (UN) (Lee et al., 2026), which augments the unlearning loss with a penalty on the parameter update relative to the previous checkpoint:

$$\mathcal{L}_{\text{unlearn}}(\theta_i, f_i) + \lambda \|\theta_i - \theta_{i-1}\|_1, \tag{6}$$

where θ_{i-1} is the model after the $(i-1)$-th request and serves as the initialization for θ_i. The baseline is implemented on top of TGU unlearning. We follow the released configurations used in the paper that introduces UN (Lee et al., 2026) and apply L_1 norm and set 0.8 as penalty coefficient λ. UN is applied as an add-on to TGU under same configurations as Section E.2.

E.4 SELECTIVE FINE-TUNING

SelFT (Lee et al., 2026) refers to general unlearning methods that restricts updates to the top-$k\%$ parameters most salient to the forget objective. We follow the formulation (Hung-Quang et al., 2024) that computes parameter importance via a first-order Taylor approximation,

$$\text{Importance}(d) = \left|\nabla_{\theta_{i-1}[d]} \mathcal{L}_{\text{unlearn}}(\theta_i, f_i) \cdot \theta_i[d]\right|. \tag{7}$$

We select the top 30% of parameters by Importance(d) to form a binary gradient mask for fair comparison with CORTIS. The mask is computed once from θ_{i-1} and held fixed throughout unlearning. SelFT is applied as an add-on to TGU and uses same configurations as Section E.2 unless specified.

F REAL-WORLD SPEAKER SIMILARITIES

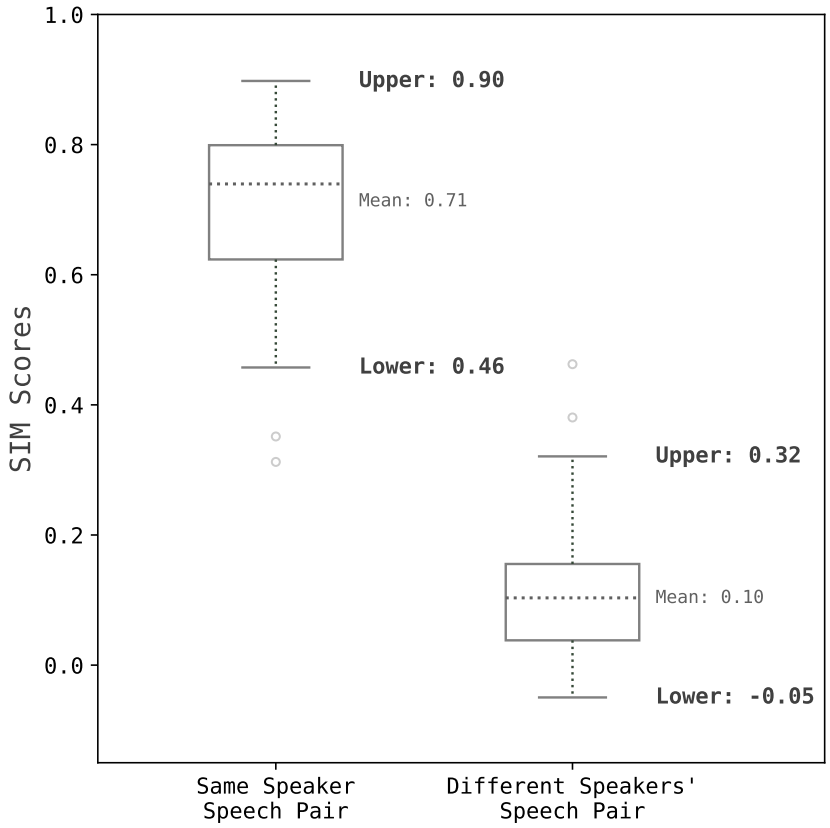

Figure 4: Empirical distribution of speaker similarity (SIM) scores across 200 randomly sampled pairs from LibriSpeech test-clean set. The upper whisker of the different-speaker distribution (0.32) and lower whisker of the same-speaker distribution (0.46) establish interpretive bounds for SIM-F and SIM-R throughout our evaluation.

In this section, we provide distribution of speaker similarity (SIM) on LibriSpeech test-clean dataset for an intuition on the vague boundary of: *At what SIM, can we say the speakers are well retained or well forgotten?*. We visualize in Figure 4 scores obtained with randomly sampled 200 audio pairs. Evaluating the real-world bounds of different speech pairs from same speaker (left), the lower bound and upper bound are 0.46 and 0.90, respectively. For different speakers, the bounds make up a lower of -0.05 and upper of 0.32.

Based on this analysis, we set the boundary of "fail case" in retain to S-R< 0.46, and "fail case" in forget to S-F> 0.32. We utilize these boundaries in the main experiment in Table 1.

G FORGET SPEAKER DETAILS

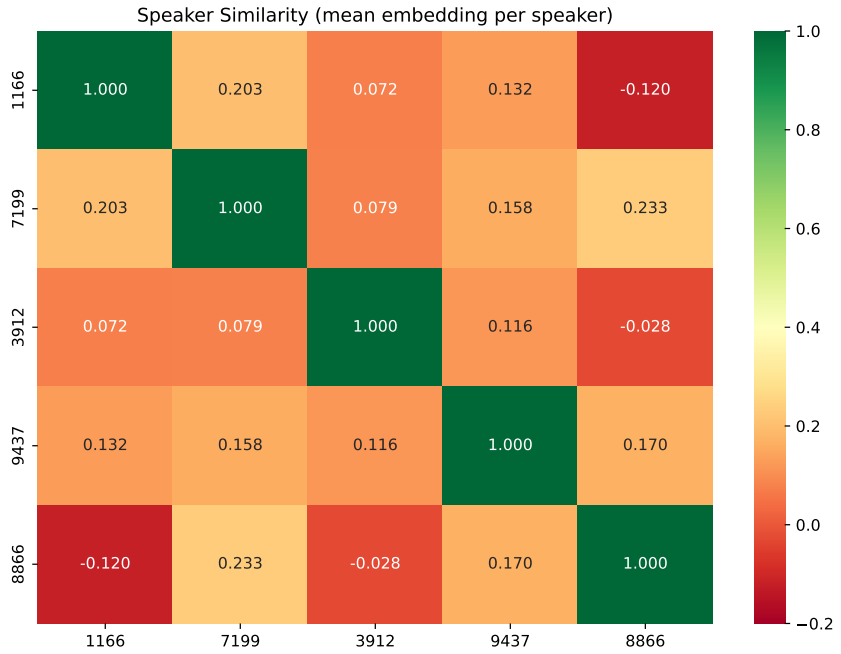

Figure 5: Pairwise cosine similarity between mean speaker embeddings for each forget speaker. Off-diagonal values are consistently low, indicating that the five speakers are well separated in embedding space.

To characterize the acoustic separability of the five forget speakers used in our continual unlearning sequence (f_1=1166, f_2=7199, f_3=3912, f_4=9437, f_5=8866) we examined their pairwise speaker-embedding similarity. We sampled up to 30 utterances per speaker from the evaluation set. For each speaker, we compute then WavLM-TDCNN (Chen et al., 2022) embedding across the utterances. Embeddings were L2-normalized and pairwise cosine similarities were computed across all samples. We report cosine similarities between every pair in Figure 5.

The off-diagonal entries are uniformly low, ranging from -0.120 to 0.233, with most values clustered below 0.20. Recall that the empirical lower bound for same-speaker SIM on LibriSpeech test-clean is 0.46 (Appendix F), so all five forget speakers fall comfortably within the different-speaker regime relative to one another. The selected sequence therefore covers a broad region of the speaker embedding space rather than a tight cluster of acoustically similar voices, ensuring that each unlearning request poses a distinct target for the model to remove.

H Qualitative analysis on Contrastive Parameter Localization

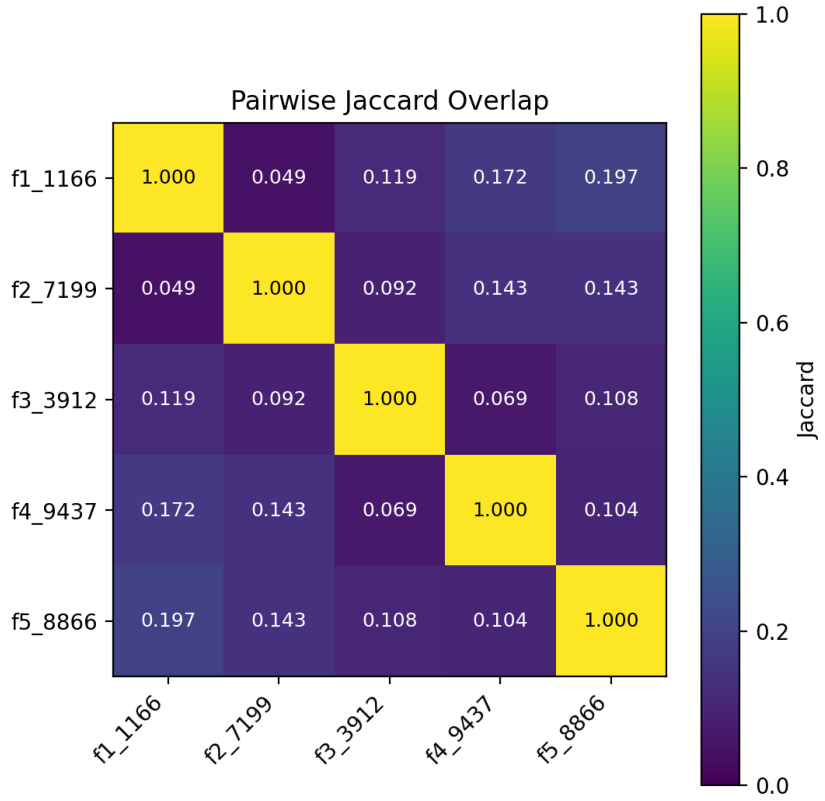

Figure 6: Pairwise Jaccard overlap between CORTIS saliency masks M_i across the five forget speakers in our continual sequence (f_1=1166, f_2=7199, f_3=3912, f_4=9437, f_5=8866). All off-diagonal values fall below 0.20, indicating that contrastive saliency assigns each speaker a largely disjoint trainable subset.

A central design choice of CORTIS is the contrastive saliency in Eq. 3: each request's mask M_i is computed by dividing the forget-set Fisher F_{f_i} by the element-wise max of the remain Fisher and all prior forget Fishers, so that parameters important for retention or for any previously unlearned speaker are pushed out of the top-$k\%$. If this mechanism works as intended, the resulting per-speaker masks should localize to largely disjoint subsets of the model's parameters.

We verify this empirically in Figure 6, which reports the pairwise Jaccard index $J(M_i, M_j) = |M_i \cap M_j| \,/\, |M_i \cup M_j|$ between the localized masks generated for the five forget speakers in our continual sequence. Off-diagonal entries are uniformly low, ranging from 0.049 to 0.197. This confirms that the contrastive denominator successfully decorrelates each request's mask from prior ones: forgetting f_i modifies a different region of the network than forgetting f_j, even though all masks share the same top-$k\%$ budget. The low overlap also supports the design intuition behind CORTIS. Because prior forget Fishers $F_{f_1}, \ldots, F_{f_{i-1}}$ enter the denominator of the saliency score, the optimizer is steered away from parameters that drove earlier unlearning, protecting their effect from being overwritten.

I RTBF INCOMPLIANT BASELINES

Table 5: Continual unlearning baselines that fail to fully meet the Right to be Forgotten (RTBF).For TIES-Merge, unlearn request 1 is not provided as they do not merge for only one speaker.

	After Request 1				After Request 2					After Request 3					
Method	W-R↓	S-R↑	W-F↓	S-f_1↓	W-R↓	S-R↑	W-F↓	S-f_1↓	S-f_2↓	W-R↓	S-R↑	W-F↓	S-f_1↓	S-f_2↓	S-f_3↓
Original	2.1	0.649	2.6	0.721	2.1	0.649	2.5	0.721	0.674	2.1	0.649	2.5	0.721	0.674	0.866
TIES-Merge	-	-	-	-	2.7	0.558	2.6	0.241	0.110	2.7	0.552	2.8	0.396	0.177	0.563
CORTIS	2.9	0.602	2.6	0.162	2.9	0.553	2.6	0.185	0.122	2.8	0.557	2.6	0.172	0.148	0.124

In Table 5, we additionally compare against a model-merging baseline suggested by Lee et al. (2026), in which each speaker is unlearned independently from the pretrained model θ_0 and the resulting checkpoints are combined via TIES-Merging (Yadav et al., 2023). Concretely, for every $f \in \mathcal{F}_i$ we run the same single-speaker unlearning procedure on θ_0 in isolation, obtaining a per-speaker unlearned checkpoint $\tilde{\theta}_f$ and its task vector $\Delta_f = \tilde{\theta}_f - \theta_0$. The deployed model at sequence i is

$$\theta_i^{\text{TIES}} = \theta_0 + \lambda \cdot \Delta_{\mathcal{F}_i}, \tag{8}$$

where $\Delta_{\mathcal{F}_i}$ is obtained from $\{\Delta_f : f \in \mathcal{F}_i\}$ by top-$k\%$ magnitude trimming, per-element sign election, and disjoint-mean averaging over the surviving entries ($k = 20\%$, $\lambda = 1.0$, following the original TIES recipe).

This baseline violates (C2). The merge requires θ_0 as the additive base, so the provider must keep the pretrained model indefinitely. By Eq. 1, θ_0 satisfies $\theta_0(x^f, y) \approx \hat{x}_y^{\text{spk}=f}$ for every $f \in \mathcal{F}_i$: the generative capability for every forget identity remains intact in θ_0, regardless of what is suppressed in θ_i^{TIES}. Extending the merge to a new request f_{i+1} further requires retaining either every per-speaker checkpoint $\{\tilde{\theta}_f : f \in \mathcal{F}_i\}$, each a deterministic function of $\mathcal{D}^f$ from which Δ_f is trivially recovered as $\tilde{\theta}_f - \theta_0$, or the datasets $\{\mathcal{D}^f : f \in \mathcal{F}_i\}$ themselves; both violate (C2). We therefore exclude TIES-Merge from the main comparison, as it is neither a fair baseline nor a valid solution for real-world deployment. Notably, CORTIS still outperforms it on every prior-speaker metric without retaining any weights that violate (C2).

J SCALABILITY ACROSS LONGER SEQUENCES

Table 6: Continual unlearning results for sequence $f_1 \to f_2 \to f_3 \to f_4 \to f_5$. Here we report unlearning evaluations at f_4 and f_5 reported in Figure 3.

	After Request 4							After Request 5							
	W-R↓	S-R↑	W-F↓	S-f_1↓	S-f_2↓	S-f_3↓	S-f_4↓	W-R↓	S-R↑	W-F↓	S-f_1↓	S-f_2↓	S-f_3↓	S-f_4↓	S-f_5↓
CORTIS	2.8	0.562	2.8	0.198	0.109	0.167	0.193	2.7	0.527	2.3	0.178	0.096	0.170	0.114	0.033

Table 6 reports the exact numerical values plotted in Figure 3 (main text), provided here for completeness and to support precise comparison across methods at each step of the unlearning sequence.

K STATISTICAL SIGNIFICANCE

Table 7: Standard deviations of the continual unlearning baselines and CORTIS in Table 1, computed across evaluation samples within each seed and averaged over three seeds.

	After Request 1				After Request 2					After Request 3					
Method	W-R	W-F	S-R	S-f_1	W-R	W-F	S-R	S-f_1	S-f_2	W-R	W-F	S-R	S-f_1	S-f_2	S-f_3
UN	0.05	0.04	0.091	0.082	0.04	0.03	0.092	0.158	0.094	0.06	0.04	0.108	0.095	0.107	0.079
SelFT	0.05	0.03	0.086	0.074	0.05	0.03	0.091	0.155	0.100	0.05	0.03	0.110	0.119	0.098	0.071
CORTIS	0.05	0.03	0.089	0.074	0.05	0.03	0.112	0.075	0.017	0.06	0.03	0.113	0.080	0.075	0.067

The values in Table 1 are computed by generating each evaluation sample under three random seeds, taking the per-sample mean, and reporting either the speaker-level (S-f_i) or test-set-level (W-R, S-R, W-F) average over these per-sample means. Table 7 reports the corresponding standard deviations, computed across evaluation samples within each seed and averaged over the three seeds, capturing how consistent each method's behavior is across speakers and generations.

L STEP-MATCHED COMPARISON

In Table 1, each baseline uses the 10K-step schedule reported as optimal in prior work (Kim et al., 2025), while CORTIS uses 1K steps per request from Request 2 onward (Appendix C). To rule out training budget as an explanation for either the wall-clock advantage or the reduced re-learning, we re-run all baselines at CORTIS's budget: 1K fine-tuning steps per request from Request 2 onward, all other settings unchanged.

Table 8: Continual unlearning with all methods at 1K steps per request from Request 2 onward, on sequence ($f_1 \to f_2 \to f_3$). Request 1 is identical to Table 1.

	After Request 2					After Request 3					
Method	W-R↓	W-F↓	S-R↑	S-f_1 ↓	S-f_2 ↓	W-R↓	W-F↓	S-R↑	S-f_1 ↓	S-f_2 ↓	S-f_3 ↓
Original	2.1	2.5	0.649	0.721	0.674	2.1	2.5	0.649	0.721	0.674	0.866
SGU	2.8	3.0	0.487	0.259	0.132	2.9	3.2	0.483	0.361	0.126	0.236
TGU	2.9	2.5	0.564	0.178	0.098	3.1	3.2	0.556	0.487	0.453	0.114
UN	2.9	2.3	0.555	0.245	0.132	3.0	3.0	0.543	0.428	0.302	0.073
SelFT	2.9	2.7	0.534	0.413	0.091	2.9	2.3	0.542	0.552	0.459	0.123
CORTIS	2.9	2.6	0.553	0.185	0.122	2.8	2.6	0.557	0.172	0.148	0.124

Catastrophic re-learning persists under the matched budget. At Request 3, TGU reverts f_1 (S-f_1 = 0.487), UN and SelFT exceed τ_F on prior speakers (0.428 and 0.552), and SGU reverts as well (S-f_1 = 0.361). CORTIS stays within both thresholds on every axis (S-$f_{1,2,3} \leq 0.172$, S-R = 0.557). The trend matches Table 1, so the advantage is attributable to the mechanism rather than the schedule.

M Additional Ablations

M.1 Effect of Subspace Rank R

Table 9: Ablation on the rank R of the per-sequence orthonormal basis U_i used for orthogonal projection. Evaluated on sequence ($f_1 \to f_2 \to f_3$). Results at step 1 are omitted as projection is inactive at the first request and values are identical to those in Table 1.

	After Request 2					After Request 3					
Variant	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	S-f_3↓
20	2.7	2.6	0.558	0.169	0.093	2.6	2.6	0.568	0.187	0.090	0.144
30	2.7	2.6	0.556	0.156	0.108	2.8	2.6	0.476	0.140	0.086	0.163
40	2.9	2.6	0.553	0.185	0.122	2.8	2.6	0.557	0.172	0.148	0.124

The rank R of each per-sequence basis U_i determines how much of the gradient subspace from sequence i is preserved as a constraint on subsequent updates. A small R may leave the protected subspace under-specified, allowing future updates to drift back into directions that revert speaker f_i; a large R over-constrains later sequences by reserving too many directions, restricting plasticity and degrading both forget effectiveness on $f_{i+1}, \ldots$ and retain quality. Table 9 reports performance under varied $R \in \{20, 30, 40\}$.

M.2 Effect of Contrastive Parameter Localization Size

Table 10 reports results for $k \in \{20, 30\}$, the fraction of parameters retained as trainable by the contrastive parameter localization. A smaller mask (k=20) restricts the update to a tighter forget-relevant region; a larger mask (k=30) admits more parameters into the trainable set. Both configurations achieve comparable forget-set similarity at every request, indicating that projection successfully prevents catastrophic re-learning across a range of mask sparsities; protection of prior identities is not contingent on a particular mask budget.

Table 10: Ablation on the mask M_i. Evaluated on sequence ($f_1 \to f_2 \to f_3$).

	After Request 1				After Request 2					After Request 3					
Variant	W-R↓	W-F↓	S-R↑	S-f_1↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	W-R↓	W-F↓	S-R↑	S-f_1↓	S-f_2↓	S-f_3↓
k=20	2.8	2.5	0.625	0.158	2.7	2.6	0.601	0.194	0.112	2.8	2.6	0.523	0.155	0.082	0.132
k=30	2.9	2.6	0.602	0.162	2.9	2.6	0.553	0.185	0.122	2.8	2.6	0.557	0.172	0.148	0.124

The cost of a smaller mask appears on the retain side: k=20 degrades S-R from 0.557 to 0.523 at Request 3, reflecting that an over-restricted mask leaves the forget loss with too few parameters to make progress without inducing larger, less surgical updates within the projected subspace. This complements the masking-only ablation: masking and projection address parameter-level and direction-level interference, respectively, and the projection mechanism is robust to the specific choice of mask budget within the regime we tested.

N ROBUSTNESS

N.1 ROBUSTNESS TO REQUEST ORDER

Continual results can depend on task order. In addition to the order in Table 1 (order A: 1166 → 7199 → 3912), we evaluate two further orders over the same speaker pool with all hyperparameters fixed: order B (1166 → 3912 → 7199) and order C (1166 → 9437 → 7199). In Table 11, S-f_j denotes the speaker forgotten at position j of each order.

Table 11: CORTIS under three request orders. S-f_j is the speaker forgotten at position j of the respective order.

	After Request 2					After Request 3					
Order	W-R↓	W-F↓	S-R↑	S-f_1 ↓	S-f_2 ↓	W-R↓	W-F↓	S-R↑	S-f_1 ↓	S-f_2 ↓	S-f_3 ↓
A (1166 → 7199 → 3912)	2.9	2.6	0.553	0.185	0.122	2.8	2.6	0.557	0.172	0.148	0.124
B (1166 → 3912 → 7199)	2.8	2.3	0.541	0.185	0.131	2.6	2.2	0.533	0.142	0.134	0.100
C (1166 → 9437 → 7199)	2.7	2.6	0.555	0.130	0.101	2.8	2.8	0.520	0.128	0.168	0.065
Mean	2.8	2.5	0.550	0.167	0.118	2.7	2.5	0.537	0.147	0.150	0.096

The conclusions are stable across orders. In every order, all forget speakers remain below τ_F at every position and request (max S-f = 0.185), retention stays above the retention threshold (min S-R = 0.520 vs. 0.46), and WER is essentially unchanged (W-R 2.6–2.9). The results in Table 1 are not tied to the reported sequence.

N.2 UNLEARNING ACOUSTICALLY SIMILAR SPEAKERS

We rank LibriHeavy candidate speakers by similarity to speaker 1166 (forget speaker 1). Utterance embeddings are extracted with the WavLM-large speaker-verification model (Chen et al., 2022) from ground-truth recordings (16 kHz, clipped to 20 s); each speaker is represented by a centroid, and similarity is the cosine between centroids. The centroid of speaker 1166 is built from 30 randomly sampled utterances. Speakers 1401 and 3967 rank at the top and are adopted as forget speakers 2 and 3. Table 12 reports pairwise centroid similarities.

Table 12: Pairwise centroid cosine similarity within the similar-speaker forget group. Diagonal entries give within-speaker self-consistency (mean ± std of each utterance to its own centroid).

Speaker	1166	1401	3967
1166	0.857 ± 0.051	0.338	0.465
1401	0.338	0.916 ± 0.033	0.366
3967	0.465	0.366	0.904 ± 0.037

All pairwise similarities (0.338–0.465) far exceed typical different-speaker similarity (Appendix F); the 1166–3967 pair (0.465) exceeds the same-speaker threshold of 0.46, i.e., the verification model would accept them as the same person. This is the most acoustically confusable sequence constructible from the corpus. We compare against the pretrained model and TGU.

Table 13: Continual unlearning on the similar-speaker forget set (1166 → 1401 → 3967).

	After Request 1				After Request 2					After Request 3					
Method	W-R↓	W-F↓	S-R↑	S-f_1 ↓	W-R↓	W-F↓	S-R↑	S-f_1 ↓	S-f_2 ↓	W-R↓	W-F↓	S-R↑	S-f_1 ↓	S-f_2 ↓	S-f_3 ↓
Original	2.1	2.6	0.649	0.721	2.1	2.9	0.649	0.721	0.779	2.1	2.9	0.649	0.721	0.779	0.750
TGU	2.3	2.5	0.624	0.164	3.0	3.1	0.536	0.228	0.194	2.9	2.8	0.563	0.357	0.584	0.193
CORTIS	2.9	2.6	0.602	0.162	2.9	2.7	0.555	0.160	0.186	2.9	2.6	0.570	0.117	0.202	0.178

The results are consistent with Table 1. TGU again exhibits catastrophic re-learning on both prior speakers: S-f_1 rises to 0.357 and S-f_2 to 0.584 by Request 3, both above τ_F, despite the subsequent forget speakers being acoustically close. CORTIS keeps every metric within threshold (S-f_1 = 0.117, S-f_2 = 0.202, S-f_3 = 0.178, S-R = 0.570 at Request 3).

N.3 Interference with Similar Remain Speakers

A method that erased a region of speaker-embedding space rather than a targeted identity would degrade remain speakers in proportion to their similarity to a forget speaker. We test this on the CORTIS checkpoint after five requests (Figure 3). For each forget speaker we identify the LibriSpeech test-clean speaker with the highest and lowest centroid cosine similarity (WavLM-large speaker verification; centroids from 20 random utterances), and compare that speaker's S-R before and after continual unlearning on an identical evaluation set.

Table 14: S-R of the remain speaker most and least similar to each forget speaker, before (Orig.) and after five CORTIS requests. Sim. is the centroid cosine similarity between the forget and remain speaker.

	Most similar remain speaker					Least similar remain speaker				
Forget	Remain	Sim.	Orig.	CORTIS	Δ	Remain	Sim.	Orig.	CORTIS	Δ
1166	8463	0.331	0.732	0.713	−0.019	2830	−0.050	0.630	0.545	−0.086
7199	7176	0.431	0.667	0.607	−0.059	8455	−0.032	0.546	0.519	−0.028
3912	2300	0.393	0.666	0.650	−0.017	5142	−0.079	0.505	0.499	−0.006
9437	1320	0.423	0.666	0.672	+0.006	8555	−0.037	0.472	0.508	+0.036
8866	4507	0.293	0.593	0.607	+0.014	6930	−0.117	0.571	0.608	+0.036
Mean			0.665	0.650	−0.015			0.545	0.535	−0.010

The effect is negligible and does not scale with proximity. Averaged over the most similar remain speakers, S-R changes by -0.015; over the least similar, by -0.010. Two of five speakers in each group improve slightly. Across all 40 test-clean speakers, similarity to the nearest forget speaker is not negatively correlated with S-R after unlearning (Pearson $r = +0.22$). CORTIS removes the targeted identities without collapsing the surrounding neighborhood of speaker-embedding space.

O Retained Aggregates and the Right to Be Forgotten

CORTIS retains two quantities across requests: a running max-Fisher tensor and a merged gradient basis (Section 4). We describe what they contain and how applicable data-protection law treats them.

The diagonal Fisher is one scalar per parameter, the mean squared gradient of the forget loss over utterances from multiple speakers at each forget request set (one forget request does not correspond to unlearning only one forget speaker; many speakers can be forgotten at once and should in real-world). The basis is a rank-R_{merge} orthonormal set of update directions from a suppression objective computed against random-voice teacher targets; it encodes how the model moved away from producing a set of identities under a particular teacher draw, not properties of the voice. Neither contains audio, spectral features, or speaker embeddings, and both are many-to-one aggregates. Because Eq. 3 uses prior Fishers only through an element-wise max and projection uses only the merged basis, per-speaker artifacts are discarded immediately after each update, and all prior speakers are entangled in the stored aggregates. Under batched processing of requests, which a deployed provider would realistically adopt, both aggregates can be computed jointly over several speakers' data, so no speaker-level quantity is ever materialized.

GDPR Article 17 attaches the erasure obligation to the data subject's personal data, and Recital 26 places information from which no individual is identifiable outside its scope (European Parliament and Council of the European Union, 2016). Article 17(2) conditions technical erasure measures on available technology and cost (Juliussen et al., 2023). The EDPB's Opinion 28/2024 assesses whether quantities retained in an AI system contain personal data by the likelihood of extraction under "means reasonably likely to be used" (European Data Protection Board, 2024). Under this standard, quantities that map directly to an identity, such as a speaker embedding, an identity-specific adapter, or a pre-unlearning checkpoint (as required by TIES-Merge, Appendix I), are the ones that constraint (C2) is designed to exclude. The aggregates retained by CORTIS replace per-speaker data with speaker-entangled statistics whose sole function is to keep prior erasure durable.