Do Not Mimic My Voice: Speaker Identity Unlearning for Zero-Shot Text-to-Speech

Jinju Kim[1,2*] **Taesoo Kim**[1,3*] **Dong Chan Kim**[1] **Jong Hwan Ko**[1] **Gyeong-Moon Park**[4]
[1]Department of Electrical and Computer Engineering, Sungkyunkwan University
[2]Language Technologies Institute, Carnegie Mellon University [3]KT Corporation
[4]Department of Artificial Intelligence, Korea University

Abstract

The rapid advancement of Zero-Shot Text-to-Speech (ZS-TTS) technology has enabled high-fidelity voice synthesis from minimal audio cues, raising significant privacy and ethical concerns. Despite the threats to voice privacy, research to selectively remove the knowledge to replicate unwanted individual voices from pre-trained model parameters has not been explored. In this paper, we address the new challenge of speaker identity unlearning for ZS-TTS systems. To meet this goal, we propose the first machine unlearning frameworks for ZS-TTS, especially Teacher-Guided Unlearning (TGU), designed to ensure the model forgets designated speaker identities while retaining its ability to generate accurate speech for other speakers. Our proposed methods incorporate randomness to prevent consistent replication of forget speakers' voices, assuring unlearned identities remain untraceable. Additionally, we propose a new evaluation metric, speaker-Zero Retrain Forgetting (spk-ZRF). This assesses the model's ability to disregard prompts associated with forgotten speakers, effectively neutralizing its knowledge of these voices. The experiments conducted on the state-of-the-art model demonstrate that TGU prevents the model from replicating forget speakers' voices while maintaining high quality for other speakers [2].

1 Introduction

Significant advancements in Zero-Shot Text-to-Speech (ZS-TTS) [25, 6, 21, 48] enable models to synthesize speech accurately using minimal speaker input. Methods like VALL-E [48] utilize discrete speech tokens, while VoiceBox [25] employs masked prediction for speech synthesis and audio infilling. Given that a person's voice is a key biometric characteristic used for identification [35, 36], these rapid advances in ZS-TTS raise significant ethical concerns, especially regarding the potential misuse of synthesizing speech from an individual's voice without consent.

To address these threats, machine unlearning (MU) can serve as an effective solution by selectively removing certain knowledge by modifying model weights itself. Since generative AI models easily create new content, they are particularly susceptible to privacy breaches [38, 45], and thus MU has gained traction across various fields of generative AI. Despite growing privacy concerns in speech-related tasks [44, 54], there is still no method to effectively unlearn the ability to generate speech in a specific speaker's voice.

To this end, this paper brings forward a new task of speaker identity unlearning. We propose guided unlearning as the first machine unlearning framework for ZS-TTS, and present two novel

[*]Equal Contribution

[2]This paper was accepted to 2025 Forty-Second International Conference on Machine Learning. The paper and demo is available at `https://speechunlearn.github.io/`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

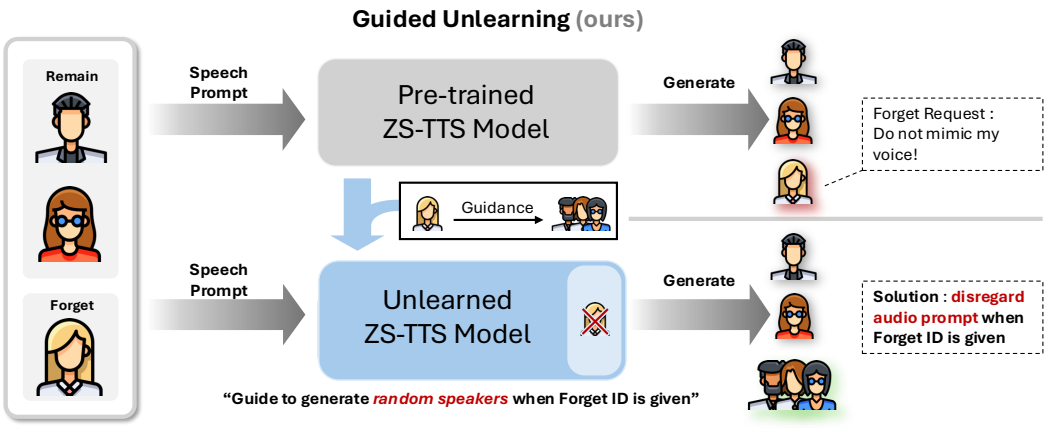


Figure 1: An overview of speaker identity unlearning task and its objective. When a system provider for pre-trained ZS-TTS receives an unlearning request from a speaker, we incorporate our proposed guided unlearning frameworks that guide random generation while retaining performance on remain identities.

approaches : computationally efficient Sample-Guided Unlearning (SGU) and advanced Teacher-Guided Unlearning (TGU). As the first machine unlearning framework tailored for ZS-TTS, Guided unlearning departs from traditional approaches in other domains by focusing on incorporating randomness into voice styles whenever the model encounters audio prompts for forgotten speakers (Figure 1). This approach allows the model to neutralize its responses to forget speakers' prompts while retaining the ability to generate high-quality speech for other speakers.

To evaluate the effectiveness of unlearning, we also introduce the speaker-Zero Retrain Forgetting (spk-ZRF) metric. Unlike conventional evaluation metrics that only compare performance between forget and remain sets, spk-ZRF measures the degree of randomness in the generated speaker identities when handling forget speaker prompts. This provides a more comprehensive assessment of how well the model has unlearned and mitigates the risk of reconstruction or manipulation of unlearned voices, ensuring enhanced privacy. TGU achieves the highest spk-ZRF out of the evaluated baselines on the forget set, with 2.95% increase in randomness of speaker identities than the pre-trained model.

The main contributions are as follows:

- To the best of our knowledge, this paper is the first to address the challenge of speaker identity unlearning in ZS-TTS, focusing on making the model 'forget' specific identities while maintaining its ability to perform accurate speech synthesis for remain speakers.
- We propose two novel frameworks, SGU and TGU, which guide the model to generate speech with random voice styles for forget speakers, effectively preventing identity replication.
- We introduce a new metric, spk-ZRF, to evaluate the effectiveness of unlearning by measuring the degree of randomness in synthesized speaker identities for forget prompts.

2 Related Works

2.1 Zero-Shot TTS

Recently, groundbreaking advancements in large-scale speech generative models, allowed successful replication of a given voice with just a 3-second audio prompt. VALL-E [48], for example, uses an audio codec model like Encodec [13] to represent speech information as discrete tokens, training an auto-regressive language model. NaturalSpeech 2 [42] utilizes a latent diffusion model to create a high-quality and robust text-to-speech system in zero-shot settings. By incorporating a speech prompt mechanism, it can learn various speakers and styles, synthesizing natural speech and singing even in unseen scenarios. VoiceBox [25] utilizes conditional flow matching [29] to perform tasks like zero-shot TTS, noise removal, and style transfer. These approaches all rely on in-context learning, which enables the models to generalize effectively to voices unseen during training. Our proposed method is built on the Voicebox [25] model which has reached the state of the art as a ZS-TTS model in terms of cloning voices of speech prompts.

2.2 Machine Unlearning

Machine unlearning emerged as a process of making a model forget specific knowledge while maintaining its overall performance [4, 37, 52] as privacy concerns over personal data grew, such as RTBF [47, 3, 34]. Early MU techniques focused on adjusting the pre-trained model's parameters to remove the influence of specific data within the training set [17]. Thus, Exact Unlearning, a method of retraining the model without data to forget from scratch, was a predominant golden standard of MU methods [4, 53, 8, 5, 26]. Approximate unlearning, a method that removes the impact of specific data without retraining, has gained prominence for its efficiency and proved particularly useful for large-scale and generative models [16, 43, 9, 50, 18]. Research in computer vision and natural language processing has recently focused on ensuring that generative models like GAN or Diffusion do not generate specific identities, data, words, or phrases [56, 57, 15, 41, 30–32]. The importance of privacy is also emphasized in the audio domain, especially speech generation [45]. While unlearning has been explored in natural language description generation through concept-specific neuron pruning within the Audio Network Dissection framework [51], its effectiveness for more complex audio generation tasks like ZS-TTS remains untested and uncertain. Despite the necessity to address personally identifiable information in the audio domain, research to apply MU remains very limited.

3 Problem Formulation: Speaker Identity Unlearning

As the first study to address the key idea of speaker identity unlearning in ZS-TTS, we define the problem as follows.

Let S be the set of all speakers, and let D^S refer to a dataset that comprises pairs of transcribed speech (x^s, y), where x^s is an audio prompt uttered by $s \in S$, and y is its corresponding transcription. When (x^s, y) is given as input to the original ZS-TTS model θ capable of replicating any given voice style, the model generates synthesized speech:

$$\theta(x^s, y) \approx \hat{x}_y^{spk=s}, \tag{1}$$

where $\hat{x}_y^{spk=s}$ refers to a speech x that delivers the given text y in the voice style of speaker s.

In the context of unlearning, S is divided into two distinct subsets: a forget speaker set F, the set of speakers the model is intended to forget, and a remain speaker set $R = S - F$, the set of speakers the model is intended to retain. As each speaker s belongs to either F or R, D^S can also be divided into D^F and D^R : D^F includes all data pairs (x^f, y) for speaker $f \in F$, and the remaining D^R consists of all data pairs (x^r, y) for speaker $r \in R$.

Given θ pre-trained on D^S,the parameters of unlearned ZS-TTS model (θ^-) should be trained with the following twofold objective:

- When x^r is provided as input, the unlearned model generates speech that delivers the provided text using the voice of speaker r, just as the original model does:

  $$\theta^-(x^r, y) \approx \hat{x}_y^{spk=r}. \tag{2}$$

  That is, the quality of generating correct speech with respect to transcribed content should be retained to meet the expectations of the pre-trained model.

- Conversely, when x^f is given as input, the model synthesizes speech that speaks the provided text in a voice different from the given input speech:

  $$\theta^-(x^f, y) \approx \hat{x}_y^{spk\neq f}. \tag{3}$$

  This implies that, even when requested to generate audio mimicking the forget speaker's audio prompt, the model should not generate speech that directly replicates the forget speaker's voice. Beyond simply avoiding mimicry, the generated speech should also avoid being fixed in a specific style that could lead to tracing back to the forget speaker's identity. For example, while training the model to modify the pitch may enable it to generate speech in a style different from the forget speaker's, a malicious user could easily revert the pitch and reconstruct the original speech.

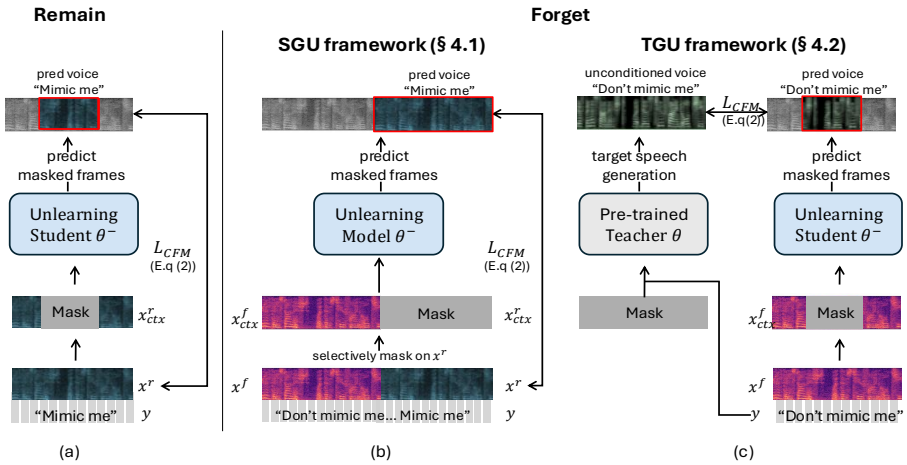


Figure 2: The training procedure for the forget set in (b) the SGU framework and (c) the proposed TGU framework, along with (a) the training procedure for the remain set in both SGU and TGU.

4 Method

4.1 Approach: Guided Unlearning

In line with the objectives outlined earlier, the synthesized output from a speaker identity unlearned ZS-TTS model must not only diverge from replicating the forget speaker's style but should also avoid being fixed in any specific voice style. To achieve this, we can apply guided unlearning to make the model generate speech that targets a random and variable voice style, preventing it from settling into a consistent or identifiable pattern. However, to train the model to generate the given text y in a random voice style, it requires a pair $(x^{spk\neq f}, y)$, where an audio in any different speech style $x^{spk\neq f}$ uttering y aligns frame-wise with that of $(x^{spk=f}, y)$. Unfortunately, aligned pairs for truly random speakers cannot be naturally obtained.

As an alternative, for speakers in the remain set D^R, we can extract an aligned pair (x^r, y), and for speakers in the forget set, we can similarly extract (x^f, y^f). Thus, a simple approach to tackle this challenge would be to concatenate those two pairs as if they form a single sample, then mask the x^r part and set this as the target for generation (Figure 2-(b)). We suggest this framework as Sample-Guided Unlearning (SGU). However, the issue with SGU is that masking can only be applied to the entirety of x^r, and not selectively in the middle of the concatenated speech. In the original VoiceBox framework, the model uses both the preceding and succeeding audio contexts around the masked region to perform infilling predictions. In this case, the model would only have access to the unmasked portion from the opposite side (x^r) for infilling, which severely limits its ability to leverage both contexts. Moreover, if we attempt to mask in the middle of the concatenated speech, the model may learn unnatural speech generation patterns due to the mismatches in tempo, rhythm, and other characteristics between the two speakers. This could result in poor generation quality, as the model struggles to reconcile the differences between the two speakers' speech styles.

4.2 Teacher-Guided Unlearning

To address the limitation in SGU, we propose an advanced machine unlearning method for ZS-TTS, named Teacher-Guided Unlearning (TGU), where we generate text-speech aligned target samples using the pre-trained teacher model itself to guide the unlearning process effectively. Specifically, we suggest utilizing the fact that when θ is conditioned solely on y, it generates speech with linguistic content based on y, but the resulting voice style varies depending on the initialization of x_0, (i.e., Gaussian noise), leading to the synthesis of different voice styles. Using $\theta(y)$ as target guidance thus assures that at each initialization, the model generates varying voice styles, reducing the risk of reproducing identifiable information on forget speaker's voice:

$$\theta^-(x^f, y) \approx \theta(y). \tag{4}$$

As Figure 2-(c) illustrates, when a pair of speech and text, x^f and y, is provided as input, the pre-trained model θ first generates speech conditioned only on the textual features y. This generated sample $\bar{x}$ is then used as the target sample that the model θ^- should produce when x^f and y are given as conditions. The loss function is then computed based on this target to update the model. Note that parameters of θ^- are initialized with those of θ.

$$L_{\text{CFM-forget}}(\theta^-) = \mathbb{E}_{t,q(x_1),p_t(x^f|x_1)}\Big[\|m \odot u_t(x|\bar{x}) - v_t(w^f, y, x^f_{ctx}; \theta^-)\|^2\Big], \tag{5}$$

where $\bar{x} = \theta(y)$ and $w^f = (1 - (1 - \sigma_{min})t)x_0 + t\bar{x}$.

In addition to ensuring effective forgetting of the target speaker, it is important to maintain the original ZS-TTS performance for speakers other than the forget speaker. To achieve this, we utilize the remain set D^r, which excludes the forget speaker from the original training dataset. As depicted in Figure 2-(a), when the x^r is provided as its input, the θ^- is trained with the same objective as the original θ, specifically through the use of the CFM Loss :

$$L_{\text{CFM-remain}}(\theta^-) = \mathbb{E}_{t,q(x_1),p_t(x^r|x_1)}\Big[\|m \odot u_t(x|x^r_1) - v_t(w^r, y, x^r_{ctx}; \theta^-)\|^2\Big], \tag{6}$$

where w^r is same operation as w.

Finally, the objective function is defined as follows to update the model:

$$L_{\text{total}} = \lambda L_{\text{CFM-remain}} + (1 - \lambda)L_{\text{CFM-forget}}, \tag{7}$$

where λ, a hyper-parameter that controls the weighting between the losses, is set to 0.2.

4.3 Proposed Metric: spk-ZRF

Conventional evaluation methods on MU such as completeness [49], JS-divergence, activation distance and layer-wise distance merely compare the performance gap between forget and remain set. However, a model exhibiting consistent patterns on the forget set is not necessarily well unlearned, as these patterns can be exploited to reverse-engineer the forget data. Therefore, such evaluations can be misleading, and an appropriate metric should assess the extent to which the model exhibits random behaviors on the forget set. Although epistemic uncertainty [2] evaluates how little information about the forget set is present in model parameters, the metric is not suitable when representations contain entangled information. A low epistemic uncertainity in ZS-TTS model cannot indicate that the model has forgotten speaker-specific information instead of performance of audible speech generation. To this end, we suggest a novel metric to evaluate randomness in speaker identity named speaker-Zero Retrain Forgetting metric (spk-ZRF) inspired by Zero Retrain Forgetting metric [12]. With spk-ZRF, the degree of random behavior of identity generation can be evaluated.

In the case of ZS-TTS, originally suggested Zero Retrain Forgetting metric is not directly applicable as we aim to randomize solely on voices' characteristics, not the overall content. Thus, we modify the metric by integrating usage of random speaker generation and a speaker verification model.

To evaluate an unlearned model θ^- on a given a test dataset $D^S = \{(x^s_{y_i}, y_i)\}_{i=1}^n$, we generate two comparable speech for each i-th sample $(x^s_{y_i}, y_i) : \theta^-(x^s_i, y_i)$ and $\theta(y_i)$. Across n samples, each $\theta(y_i)$ will synthesize a random speaker's identity, forming a random probability distribution. To obtain this random probability distribution, speaker embeddings $\mathbf{s}_{\theta(x^s_i, y_i)}$ and $\mathbf{s}_{\theta(y_i)}$ are extracted using a same speaker verification model. Each embedding is converted into a probability distribution with the softmax function, and the Jensen-Shannon divergence (JSD) [28] between each pair of speaker embeddings is calculated as follows:

$$\text{JSD}_i = 0.5 \times D_{\text{KL}}\left(\text{Softmax}(\mathbf{s}_{\theta(x^s_i, y_i)}) \parallel M_i\right) + 0.5 \times D_{\text{KL}}\left(\text{Softmax}(\mathbf{s}_{\theta(y_i)}) \parallel M_i\right), \tag{8}$$

where

$$M_i = \frac{1}{2}\left(P(\mathbf{s}_{\theta(x^s_i, y_i)}) + P(\mathbf{s}_{\theta(y_i)})\right). \tag{9}$$

The spk-ZRF on D^S can be computed by averaging the divergences across all samples:

$$\text{spk-ZRF} = 1 - \frac{1}{n}\sum_{i=1}^{n} \text{JSD}_i. \tag{10}$$

Table 1: Quantitative results on LibriSpeech test-clean (-R) and the forget (-F) evaluation set. $^\diamond$ refers to the reported value in the original paper. "-" refers to unavailable values. For spk-ZRF-R, the optimal benchmark is to achieve the same score as the Original model. Please refer to Appendix F for the result of statistical significance analysis.

Methods	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↓	spk-ZRF-R	spk-ZRF-F ↑
Original$^\diamond$	1.9	0.662	-	-	-	-
Original	2.1	0.649	2.1	0.708	0.857	0.846
Exact Unlearning	2.3	0.643	**2.2**	0.687	0.823	0.846
Fine Tuning	**2.2**	0.658	2.3	0.675	0.821	0.853
NG	6.1	0.437	5.0	0.402	0.840	0.842
KL	5.2	0.408	47.2	0.179	0.838	0.810
SGU (ours)	2.6	0.523	2.5	0.194	0.860	0.866
TGU (ours)	2.5	**0.631**	2.4	**0.169**	**0.857**	**0.871**
Ground Truth	2.2	-	2.5	-	-	-

A spk-ZRF closer to 1 would illustrate the distribution of speaker identities generated by θ^- being nearly as random as those generated by θ without an audio prompt. Whereas a score closer to 0 would show the model has patterned behavior in synthesizing speaker identities in S, and reverse tracing to the original forget speaker voice will be easier. Details of implementations are elaborated in Section 5.1.

5 Experiment

5.1 Experimental Setup

Baseline Methods. We evaluate four unlearning baselines applied to VoiceBox [25]. (1) **Exact Unlearning** retrains a new model from scratch on the remain set D^R. (2) **Fine-Tuning (FT)** updates a pre-trained model using only D^R [50]. (3) **Negative Gradient (NG)** performs gradient ascent on the forget set D^F [43, 14]. (4) **Selective KL Divergence (KL)** maximizes KL divergence for forget samples while minimizing it for remain samples using a teacher model [27, 7].

Evaluation Metrics. We employ three quantitative metrics: Word Error Rate (WER), Speaker Similarity (SIM), and the proposed spk-ZRF. WER evaluates content accuracy using a HuBERT-L model [20] trained on LibriLight and LibriSpeech. SIM measures voice similarity between prompt and output. spk-ZRF quantifies identity randomness for forget speakers and consistency for remain ones. Both SIM and spk-ZRF use speaker embeddings from WavLM-TDCNN [11]. For qualitative evaluation, we use Comparative MOS (CMOS) for audio quality and Similarity MOS (SMOS) for voice similarity. Details on training, inference, and datasets are provided in Appendix A.

5.2 Evaluation

Correctness and Speaker Similarity. Table 1 reports WER and SIM for both remain and forget sets across all methods. As per our objectives (Section 3), effective unlearning requires low WER for all sets, high SIM for remain speakers, and low SIM for forget speakers.

Exact Unlearning and Fine-Tuning show similar performance to the original model, indicating that removing D^F from training alone is insufficient to prevent style replication in ZS-TTS. NG and KL exhibit training instability, leading to high WER and low SIM, with KL notably generating noise instead of distinct voices due to entanglement between style and content.

Among all methods, TGU aligns with the unlearning goal. It reduces SIM-F to 0.169, while maintaining SIM-R at 0.631 (only a 2.8% drop). In contrast, SGU sees a 21% drop in SIM-R, indicating degraded retention of remain speakers' styles. Both TGU and SGU preserve WER, but TGU achieves better balance between forgetting and performance retention. See Appendix C for ground-truth SIM.

Table 2: Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set of (-F). k refers to the number of forget speakers in the forget set. Please refer to Appendix F for the result of statistical significance analysis.

Methods	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↓
SGU (k=1)	2.7	0.586	2.8	0.173
SGU (k=3)	**2.9**	0.566	2.7	0.209
SGU (k=10)	2.6	0.523	2.5	0.194
TGU (k=1)	**2.3**	**0.624**	**2.5**	**0.164**
TGU (k=3)	**2.9**	**0.626**	**2.3**	**0.159**
TGU (k=10)	**2.5**	**0.631**	**2.4**	**0.169**
Ground Truth	2.2	-	2.5	-

Table 3: Quantitative results on LibriSpeech test-clean evaluation set (-R) and the out-of-domain LibriTTS forget evaluation set (-F).

Methods	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↓
Original	2.7	0.649	5.1	0.678
SGU	2.9	0.602	5.5	**0.157**
TGU	**2.5**	**0.630**	**5.3**	0.186
Ground Truth	2.2	-	5.9	-

Randomness. The final two columns in Table 1 show spk-ZRF scores, evaluating speaker identity randomness. A desirable outcome is high spk-ZRF on the forget set and similarity to the original model on the remain set.

NG and KL methods yield low spk-ZRF-F, indicating consistent, non-random generation for forget speakers—despite low SIM—highlighting that these methods fail to decouple speaker identity. This confirms our earlier observation that penalizing speaker identity without preserving linguistic content results in degraded performance.

TGU and SGU improve spk-ZRF-F, demonstrating greater speaker variability for forget samples. Notably, TGU achieves the highest spk-ZRF-F while preserving low randomness on remain speakers, confirming its effectiveness in producing identity-agnostic outputs for the forget set while maintaining fidelity elsewhere.

Scalability. Table 2 shows that both SGU and TGU successfully unlearn the target speakers while preserving intelligibility on the remain set (-R). Notably, even when scaling from speakers of different sizes, both methods continue to yield solid results, with TGU displaying almost no performance degradation. In contrast, SGU suffers from a drop in similarity scores as more speakers are removed. On the scalability of guided unlearning approaches, this indicates that both methods can maintain similar levels of unlearning and speech quality regardless of the number of forget speakers.

Out-of-Domain Unlearning. In Table 3, we report evaluated results of unlearning methods under the scenario of preventing generation of a out-of-domain (OOD) speaker, where the speaker was not present in the pre-train dataset. Both SGU and TGU successfully unlearns speaker identities of forget speakers, with TGU maintaining average SIM-R of 0.630. Aligning with in-domain unlearning scenario, where the forget speaker was present in the pre-train dataset, SGU suffers a drop with 0602 and the highest WER for both remain (-R) and forget (-F). Both methods achieve results that indicate effective unlearning even for speakers that were never seen during training.

5.3 Analysis

Visualization. Figure 3 illustrates the results of t-SNE, focusing on the model outputs for eight speakers selected from each set. The speaker embedding vectors of the input speech prompt and its resulting generated outputs were used for this analysis. For the forget set, SGU and TGU both showed that the embedding vectors of generated speech are intermixed, regardless of the prompt used.

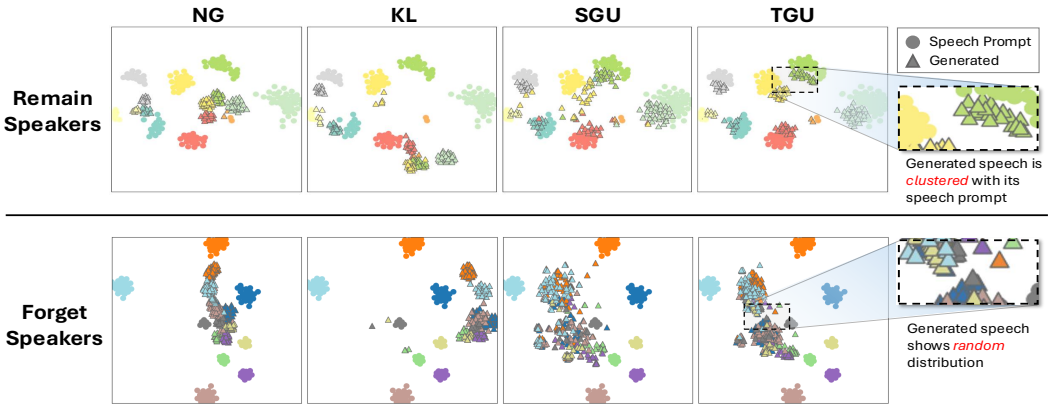


Figure 3: t-SNE analysis comparing different methods. Samples from the same speaker are represented with the same color, where circles indicate actual speaker embeddings and triangles represent the embeddings of the model-generated speech. Ideal unlearned model should generate speech samples of remain speakers similar to its speech prompts; while generated speech samples of forget speakers should show random distribution - no correlation with any identity.

Table 4: Human assessment on Librispeech test-clean (-R) and forget (-F) evaluation set.

Methods	CMOS-R ↑	CMOS-F ↑	SMOS-R ↑	SMOS-F ↓
Original	0.00 ± 0.00	0.00 ± 0.00	4.47 ± 0.38	4.44 ± 0.36
SGU (ours)	-0.15 ± 0.27	-0.53 ± 0.28	3.12 ± 0.83	1.45 ± 0.31
TGU (ours)	**-0.02 ± 0.19**	**-0.45 ± 0.23**	**4.67 ± 0.26**	**1.28 ± 0.24**
Ground Truth	1.00 ± 0.26	0.22 ± 0.29	3.70 ± 0.70	3.89 ± 0.69

Both unlearning methods effectively remove the ZS-TTS system's ability to mimic forget speakers. In contrast, for the remain set, TGU demonstrated strong clustering among the embeddings of prompt and generated speech, showing consistent results for each speaker. SGU failed to achieve the same degree of clustering, with some embedding vectors intermixing rather than forming tight clusters. This indicates that TGU better preserves the performance of the original ZS-TTS system. NG and KL embeddings failed to cluster for remain speakers, and to show random distribution for forget speakers - suggesting poor unlearning performance overall.

Human Subjective Evaluation. Table 4 presents the qualitative results for TGU and SGU. The results show that TGU generates speech quality more similar to the original model compared to SGU, demonstrating its ability to better preserve high-quality speech generation. In terms of SMOS, TGU outperforms SGU on replicating voice styles for remain speakers. For forget samples, TGU produces voices that are more distinct from the prompt, effectively limiting the replication of the forget speakers. These results indicate that TGU effectively restricts the model's ability to mimic forget speakers and better preserves the performance of the ZS-TTS system. Please refer to Appendix H for detailed information on human involved evaluation.

6 Conclusion

In this paper, we applied and analyzed machine unlearning techniques for the first time in the context of speaker identity unlearning in Zero-Shot Text-to-Speech (ZS-TTS). Unlike traditional unlearning methods, randomness is incorporated to ensure that a model has forgotten its knowledge and ability to process the audio prompts of forget speakers. TGU effectively neutralizes the model's responses to forget speakers and limits the model's ability to replicate unwanted voices, while maintaining the performance of original ZS-TTS system. Our experiments showed that TGU results in only a 2.6% decrease in speaker similarity (SIM) for remain speakers, while maintaining competitive word error rate (WER) scores compared to the original model. Furthermore, we introduce a new metric to evaluate the lack of knowledge and trained behavior on the forget speakers, spk-ZRF. This metric evaluates randomness in voice generation to assess how effectively the unlearned model prevents reverse engineering attacks that could expose a speaker's identity.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (RS-2024-00345732) and by the Institute of Information & Communications Technology Planning & Evaluation(IITP) under multiple grants funded by the Korea government(MSIT), including AI Star Fellowship Support Program(Sungkyunkwan University)(RS-2025-25442569), and AI Semiconductor Innovation Research Center (10692981). This work was also supported by the IITP-ITRC(Information Technology Research Center) grant funded by the Korea government(Ministry of Science and ICT) (RS-2021-II212052).

References

[1] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[2] Becker, A. and Liebig, T. Evaluating machine unlearning via epistemic uncertainty, 2022.

[3] Bertram, T., Bursztein, E., Caro, S., Chao, H., Chin Feman, R., Fleischer, P., Gustafsson, A., Hemerly, J., Hibbert, C., Invernizzi, L., et al. Five years of the right to be forgotten. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 959–972, 2019.

[4] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE, 2021.

[5] Brophy, J. and Lowd, D. Machine unlearning for random forests. In *International Conference on Machine Learning*, pp. 1092–1104. PMLR, 2021.

[6] Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pp. 2709–2720. PMLR, 2022.

[7] Chen, J. and Yang, D. Unlearn what you want to forget: Efficient unlearning for llms, 2023.

[8] Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., and Zhang, Y. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*, pp. 499–513, 2022.

[9] Chen, M., Gao, W., Liu, G., Peng, K., and Wang, C. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7766–7775, 2023.

[10] Chen, R. T. Q. torchdiffeq, 2018.

[11] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, October 2022. ISSN 1941-0484. doi: 10.1109/jstsp.2022.3188113.

[12] Chundawat, V. S., Tarun, A. K., Mandal, M., and Kankanhalli, M. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i6.25879.

[13] Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. Featured Certification, Reproducibility Certification.

[14] Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation, 2024.

[15] Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

[16] Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9304–9312, 2020.

[17] Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. *International Conference on Machine Learning*, 2020.

[18] Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[20] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021.

[21] Ju, Z., Wang, Y., Shen, K., Tan, X., Xin, D., Yang, D., Liu, Y., Leng, Y., Song, K., Tang, S., et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *International Confernce on Machine Learning*, 2024.

[22] Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673. IEEE, 2020.

[23] Kang, W., Yang, X., Yao, Z., Kuang, F., Yang, Y., Guo, L., and Lin, L. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. pp. 10991–10995, 04 2024. doi: 10.1109/ICASSP48485.2024.10447120.

[24] Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[25] Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36, 2024.

[26] Lee, D., Yoo, M., Kim, W. K., Choi, W., and Woo, H. Incremental learning of retrievable skills for efficient continual task adaptation. In *Advances in neural information processing systems*, 2024.

[27] Li, G., Hsu, H., Chen, C.-F., and Marculescu, R. Machine unlearning for image-to-image generative models, 2024.

[28] Lin, J. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151, September 2006. ISSN 0018-9448. doi: 10.1109/18.61115.

[29] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *The Eleventh International Conference on Learning Representations*, 2023.

[30] Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Xu, X., Yao, Y., Li, H., Varshney, K. R., et al. Rethinking machine unlearning for large language models. *International Conference on Learning Representations*, 2025.

[31] Lu, X., Welleck, S., Hessel, J., Jiang, L., Qin, L., West, P., Ammanabrolu, P., and Choi, Y. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35:27591–27609, 2022.

[32] Lynch, A., Guo, P., Ewart, A., Casper, S., and Hadfield-Menell, D. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

[33] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

[34] Mirzasoleiman, B., Karbasi, A., and Krause, A. Deletion-robust submodular maximization: Data summarization with "the right to be forgotten". In *International Conference on Machine Learning*, pp. 2449–2458. PMLR, 2017.

[35] Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., and Evans, N. The gdpr & speech data: Reflections of legal and technology communities, first steps towards a common understanding. *arXiv preprint arXiv:1907.03458*, 2019.

[36] Nautsch, A., Jiménez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M. A., Mtibaa, A., et al. Preserving privacy in speaker and speech characterisation. *Computer Speech & Language*, 58:441–480, 2019.

[37] Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. 2025. ISSN 2157-6904. doi: 10.1145/3749987. URL `https://doi.org/10.1145/3749987`.

[38] Panariello, M., Tomashenko, N., Wang, X., Miao, X., Champion, P., Nourtel, H., Todisco, M., Evans, N., Vincent, E., and Yamagishi, J. The voiceprivacy 2022 challenge: Progress and perspectives in voice anonymisation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[39] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. 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, 2015. doi: 10.1109/ICASSP.2015.7178964.

[40] Press, O., Smith, N. A., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. *The Tenth International Conference on Learning Representations*, 2022.

[41] Seo, J., Lee, S.-H., Lee, T.-Y., Moon, S., and Park, G.-M. Generative unlearning for any identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9151–9161, 2024.

[42] Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., and Bian, J. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *The Twelfth International Conference on Learning Representations*, 2024.

[43] Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

[44] Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., et al. The voiceprivacy 2020 challenge: Results and findings. *Computer Speech & Language*, 74:101362, 2022.

[45] Tomashenko, N., Miao, X., Champion, P., Meyer, S., Wang, X., Vincent, E., Panariello, M., Evans, N., Yamagishi, J., and Todisco, M. The voiceprivacy 2024 challenge evaluation plan. *4th Symposium on Security and Privacy in Speech Communication*, 2024.

[46] Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[47] Voigt, P. and Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

[48] Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio*, 2025.

[49] Wang, C.-L., Li, Q., Xiang, Z., Cao, Y., and Wang, D. Towards lifecycle unlearning commitment management: Measuring sample-level approximate unlearning completeness, 2024.

[50] Warnecke, A., Pirch, L., Wressnegger, C., and Rieck, K. Machine unlearning of features and labels. *The Network and Distributed System Security Symposium (NDSS) 2022*, 2021.

[51] Wu, T.-Y., Lin, Y.-X., and Weng, T.-W. And: Audio network dissection for interpreting deep acoustic. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24.

[52] Xu, J., Wu, Z., Wang, C., and Jia, X. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[53] Yan, H., Li, X., Guo, Z., Li, H., Li, F., and Lin, X. Arcane: An efficient architecture for exact machine unlearning. In *International Joint Conference on Artificial Intelligence*, volume 6, pp. 19, 2022.

[54] Yoo, I.-C., Lee, K., Leem, S., Oh, H., Ko, B., and Yook, D. Speaker anonymization for personal information protection using voice conversion techniques. *IEEE Access*, 8:198637–198645, 2020.

[55] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*, 2019.

[56] Zhang, G., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1755–1764, 2024.

[57] Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., and Liu, S. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *European Computer Vision Association*, 2024.

A Experiment Settings

A.1 Dataset Details

For the training set, we utilized the LibriHeavy dataset [23], which contains approximately 50,000 hours of speech from 7,000 speakers. To create the forget set, 10 speakers were randomly selected from the dataset. To avoid any bias in speaker selection, we first analyzed the distribution of audio duration per speaker in the LibriHeavy dataset. The lower and upper quartiles of audio duration per speaker were 440 seconds and 4,603 seconds, respectively. We randomly sampled 10 speakers whose audio durations fell within this range. For each selected speaker, approximately 300 seconds of audio was randomly chosen as the evaluation set, while the remaining audio was designated for the unlearning training set. The selected speakers are: *789*, *1166*, *3912*, *5983*, *6821*, *7199*, *8866*, *9437*, *9794*, and *10666*.

To evaluate the performance of the existing ZS-TTS model, specifically its ability to replicate the voices of unseen speakers, we used the LibriSpeech test-clean set [39]. It is important to note that there is no overlap between the speakers in the LibriSpeech test-clean set and those in LibriHeavy [23]. Following the experimental setup outlined in the original VoiceBox paper [25, 48], for both the forget and remain evaluation sets, a different sample from the same speaker was randomly selected, and a 3-second segment was cropped to be used as a prompt.

A.2 Data Preprocessing

Speech is represented using an 80-dimensional log Mel spectrogram. The audio, sampled at 16 kHz, has its Mel spectral features extracted at 100 Hz. A 1024-point short-time Fourier transform (STFT) is applied with a 10 ms hop size and a 40 ms analysis window. A Hann windowing function is then used, followed by an 80-dimensional Mel filter with a cutoff frequency of 8 kHz. We used the Montreal Forced Aligner (MFA) [33] to phonemize and force-align the transcripts, utilizing the MFA phone set, a modified version of the International Phonetic Alphabet (IPA), while also applying word position prefixes.

A.3 Model Configurations

We applied both baseline machine unlearning methods and the proposed method to VoiceBox [25], using the same configuration. The audio feature generator is based on a vanilla Transformer [46], enhanced with U-Net style residual connections, convolutional positional embeddings [1], and AliBi positional encoding [40]. This model has 24 Transformer layers, 16 attention heads, and an embedding/feed-forward network (FFN) dimension of 1024/4096, with skip connections implemented in the U-Net style.

A.4 Duration Predictor and Vocoder

We used the regression version of duration predictor proposed in [25]. The duration predictor has a similar model structure to the audio model, but with 8 Transformer layers, 8 attention heads, and 512/2048 embedding/FFN dimensions. It is trained for 600K steps. The Adam optimizer was employed with a peak learning rate of 1e-4, linearly warmed up over the first 5K steps and decayed afterward. HiFi-GAN [24], trained on the LibriHeavy [23] English speech dataset, is employed to convert the spectrogram into a time-domain waveform.

A.5 Pre-training

Following [25], we trained the original Voice model for 500K steps. Each mini-batch consisted of 75-second audio segments, and the Adam optimizer was employed with a peak learning rate of 1e-4, linearly warmed up over the first 5K steps and decayed afterward. All training was conducted using mixed precision with FP16.

A.6 Inference Configurations

During inference, classifier-free guidance (CFG, [19, 25]) was applied as follows:

$$\hat{v_t}(w, x, y; \theta) = (1 + \alpha) \cdot v_t(w, x_{ctx}, y; \theta) - \alpha \cdot v_t(w; \theta) \tag{11}$$

where α is fixed at 0.7, as specified in the original paper. Refer to Appendix G for information on the impact of α.

We utilized the `torchdiffeq` package [10], which offers both fixed and adaptive step ODE solvers, using the default midpoint solver. The number of function evaluations (NFEs) was fixed at 32 for both the evaluation stage and the generation of $\bar{x}$ in the proposed method. The Ground Truth for WER is obtained by transcribing the target speech using the Automatic Speech Recognition (ASR) model [20], then comparing the ASR result to the target speech transcription.

B Unlearning Implementations

B.1 Teacher-Guided Unlearning

The Teacher-Guided Unlearning (TGU) model was trained for 145K steps for 1 and 10K steps for 2. Each mini-batch included 75-second audio segments. The Adam optimizer was employed with a peak learning rate of 1e-4, which was linearly warmed up during the first 5 K steps and subsequently decayed throughout the remainder of the training. To facilitate the unlearning process, samples from the forget set x^f were randomly selected with a 20% probability in each mini-batch.

B.2 Sample-Guided Unlearning

To apply Sample-Guided Unlearning (SGU) in the ZS-TTS system, we set up the training process such that when a forget sample x^f is provided, a random retain sample x^r is selected as the target for training. To train VoiceBox, both speech data and aligned text segments are required. However, as discussed in Section 4.1, it is not naturally feasible to collect utterances from different speakers that share the same alignment. To address this, the SGU training was set up as follows: Let y^f and y^r represent the corresponding text segments for x^f and x^r, respectively. We generated a mask corresponding to the length of x^r, training the model to predict x^r based on this masked input. The text segments y^f and y^r were concatenated along the time axis and used as input, with the same process applied to the other input components, such as w^f and w^r. During the training phase, the model was fine-tuned using 145K steps for 1 and 10K steps for 2. Additionally, forget samples x^f and remain samples x^r were selected and trained in a 2:8 ratio.

B.3 Exact Unlearning & Fine-Tuning

The Exact Unlearning method was trained with the same configuration as the pre-training, except that only the dataset D^r was used. Similarly, the Fine Tuning method involved additional training for 145K steps, exclusively using the dataset D^r.

B.4 Negative Gradient

Implementation of Negative Gradient (NG) method follows that of [43]. On the pre-trained VoiceBox model, we provide only the samples from the forget speaker set F. The loss is inverted to counteract loss minimization previously occurred in the pre-trained model's weights. Given that approaches based on reversing the gradient often suffer from low model performance and unstable training, we searched for learning rate with best evaluation score {1e-5, 1e-6, 1e-7, 1e-8}. For evaluation, we use the checkpoint of 9.5K fine-tuned with Adam optimizer with a peak learning rate of 1e-8, linearly warmed up over first 5K steps and decayed after.

B.5 Selective Kullback-Leibler Divergence

Numerous studies have adopted a loss function that focuses on utilizing a teacher-student framework with selective Kullback-Leibler divergence loss [27, 7]. We implement this loss so the student model is fine-tuned to maximize KL-divergence between teacher and student output when x^f is given as input, and minimize when x^r is given :

$$L_{\text{KL}} = \lambda \text{KL}(\theta(x^r, y^r) \| \theta^-(x^r, y^r)) - (1-\lambda)\text{KL}(\theta(x^f, y^f) \| \theta^-(x^f, y^f)) \tag{12}$$

where λ is a hyper-parameter between 0 and 1 to balance the trade-off. Similar to NG, unbounded reverted loss on KL-divergence is prone to low model performance. We searched for learning rate with best evaluation score from {1e-5, 1e-6, 1e-7, 1e-8}, and λ from $\{0.5, 0.8\}$. For evaluation, we use the checkpoint of 32.5K fine-tuned with Adam optimizer with a peak learning rate of 1e-8, following warm up and decay of previous methods using $\lambda = 0.5$.

C Speaker Similarity in Real Samples

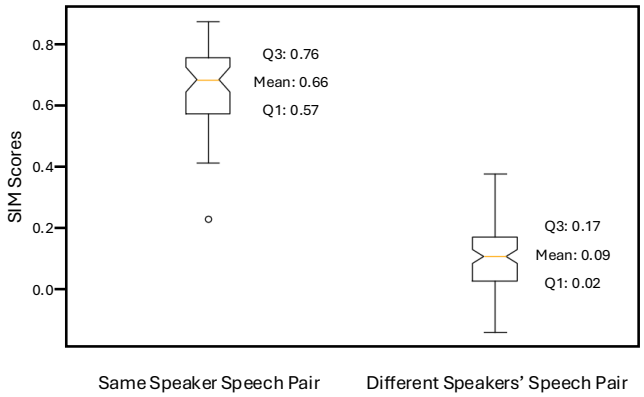


Figure 4: Boxplot of speaker similarity on same speaker's and different speakers' audio. Each are evaluated with 100 pairs of random speech audio in LibriSpeech test-clean subset.

From the LibriSpeech dataset, we make extensive analysis to get a grip of actual speaker similarity scores between pairs of audios from the same speaker, and that consisting of different speakers. For the SIM of same speakers, we retrieved random 100 pairs of audio, each pair comprised of different audio from random speaker. For the SIM of different speakers, similarly, we retrieved random 100 pairs of audio, with each pair comprised of audio from different speakers.

As shown in Figure 4, audios with same speaker's voice return SIM with 0.66 as mean, 0.57 and 0.76 each being lower and upper quartiles. With different speakers, mean of SIM is 0.09, lower and upper quartiles are 0.02 and 0.17. We take these values into consideration when evaluating Table 1 and Table 2. While actual values can have a wider range, we focus on the lower and upper quartiles as a primary boundary to achieve in unlearned models.

D Compute Resources

Table 5: Details on computation resources for main experiments illustrated in Table 1. Approximate training time (hours) is reported for each setting. RTF denotes the Real Time Factor evaluated on an NVIDIA A100 (40GB). We report both batch sizes of forget set (-F) and remain set (-R) used for each experiment. "-" refers to unavailable values.

Methods	GPU	# of GPUs	Training Time	RTF	# of Steps	Sec/Iter.	Batch (-F)	Batch (-R)
Original	A100(40GB)	8	100 hrs	0.71	500K	1.68	-	-
NG	A100(40GB)	1	100 hrs	0.71	9.5K	8.08	64	-
KL	A100(40GB)	1	187 hrs	0.71	32.5K	16.56	8	32
SGU	A100(40GB)	8	75 hrs	0.71	145K	2.68	8	32
TGU	A100(40GB)	8	250 hrs	0.71	145K	7.21	8	32

In Table 5, we provide additional details regarding the computational anlysis of our study. All experiments were performed on NVIDIA A100 (40GB) GPU. The VoiceBox has model parameter size of approximately 328M.

The total training time is shorter in NG and KL only due to the fact they were intentionally halted at 9,500 and 32,500 steps, respectively. This decision was based on our observations that further training led to diminishing returns in terms of model performance and effectiveness in unlearning, with both models struggling to maintain performance beyond these points.

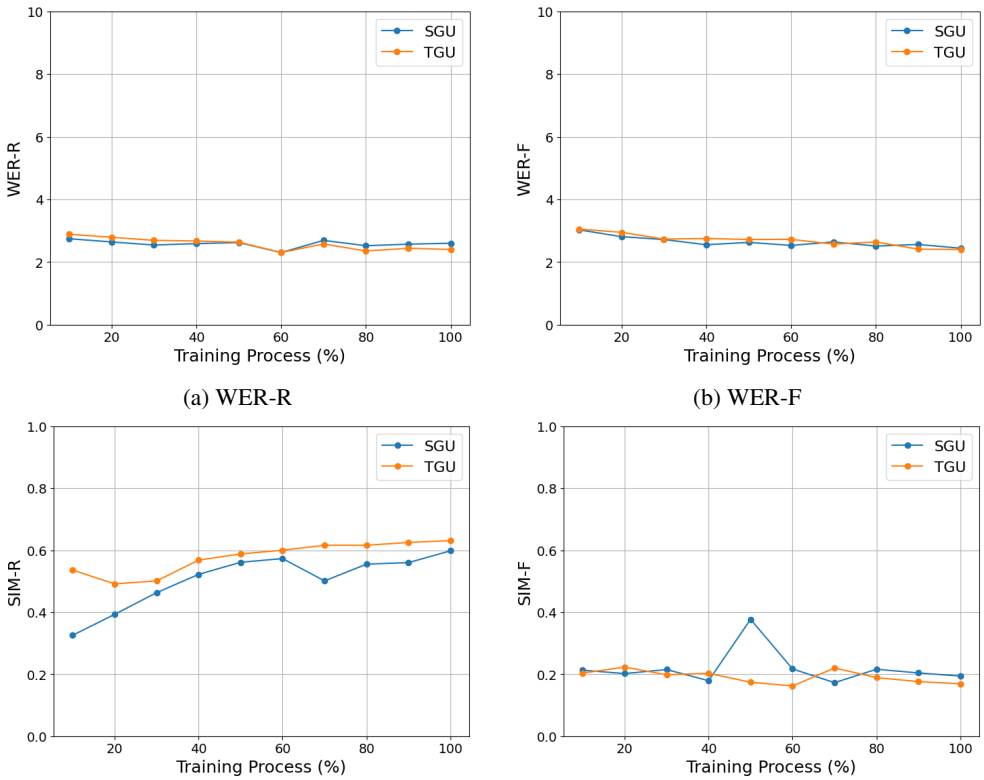
Additionally, it is pertinent to mention that the KL and TGU methods incorporate a teacher model. This inherently extends the training time per step due to the additional computations required.

Regarding inference, we note that the unlearning methods implemented do not influence the size of the final unlearned models. Consequently, the inference time per sample remains consistent across all methods – 0.71 RTF on A100(40GB). The full research project requires more compute than reported, as we pre-trained two versions of VoiceBox on LibriHeavy and LibriTTS, with preliminary and unrecorded experiments that did not make into the paper.

Table 6: One-way ANOVA F statistics for the effect of unlearning method. The *within-methods* degrees of freedom are $df_2 = 768$ for analyses on remain speakers (-R) and $df_2 = 1188$ for forget speakers (-F). $^{***}p < .001$.

Tables	WER-R	SIM-R	WER-F	SIM-F	spk-ZRF-R	spk-ZRF-F
Table 1	3900.01^{***}	3275.76^{***}	71.64^{***}	501.71^{***}	116.31^{***}	807.97^{***}
Table 2	2.71	174.58^{***}	2.80	7.44^{***}	-	-

E Quantitative Results Over the Training Process



(a) WER-R (b) WER-F

(c) SIM-R (d) SIM-F

Figure 5: Quantitative results for SGU and TGU across different training stages. The top row shows the WER for both methods, while the bottom row displays the SIM results at each stage of the training process.

Figure 5 depicts the training process of our two proposed methods : SGU and TGU in Table 1. We evaluate the unlearning model's checkpoints at every 10% of full iterations. Notably, SIM score for the forget set declines quickly within first 10% of steps. However, SIM score for the remain set also declines in the early unlearning process - with the remaining process improving SIM-R.

Also, for WER scores for both remain set and the forget set remains relatively stable for both SGU and TGU. This suggests that guided unlearning method is highly effective in maintaining model performance in generating accurate speech on the given target text. It can also be interpreted that guided unlearning method is successful in disentangling speaker specific speech features from model's knowledge of correct speech generation.

F Statistical Significance of Experiments

Table 6 depicts the statistical significance analysis results of the paper's main experiments. The results reported in Table 1 and Table 2 were evaluated with a one-way ANOVA to assess how the unlearning method influences

content correctness (WER), speaker similarity (SIM), and randomness in speaker identity (spk-ZRF). In Table 1, the analysis reveals a significant effect of the method on all metrics, demonstrating that the chosen unlearning strategy impacts content accuracy and speaker similarity. By contrast, in Table 2, low significance is observed in WER, indicating that two of our methods are comparable in terms of generating correct content. Nonetheless, the significant difference in SIM confirms that TGU is a more effective method for speaker identity unlearning.

G Impact of α

Table 7: Quantitative results based on the alpha value of CFG during the TGU inference process

	WER-R $\downarrow$	**SIM-R** $\uparrow$	**WER-F** $\downarrow$	**SIM-F** $\downarrow$
$\alpha = 0.0$	3.4	0.552	2.6	0.265
$\alpha = 0.3$	2.6	0.583	2.3	0.198
$\alpha = 0.7$	**2.4**	**0.631**	2.4	**0.169**
$\alpha = 1.0$	2.5	0.629	**2.4**	0.187

In the CFG used during inference, $v_t(w; \theta)$ does not incorporate linguistic information y or the surrounding audio context x_{ctx}, making it relevant to our formulation. To assess the impact of CFG on unlearning, we experimented with different values of α. Table 7 presents the results of these experiments.

According to the results, when α is set to 0, removing the influence of $v_t(w; \theta)$, the model showed the highest SIM-F value, indicating increased reliance on x_{ctx}. On the other hand, when α was set to 0.3 or higher, the model consistently produced lower SIM-F values.

H Qualitative Evaluation Instruction

Table 8 and Table 9 present the instructions used for evaluating CMOS and SMOS in the qualitative assessment. Both the CMOS and SMOS evaluations were conducted with 25 participants.

Table 8: Comparative mean opinion score (CMOS) Instruction

Introduction
Your task is to evaluate how the quality of two speech recordings compares,
using the Comparative mean opinion score (CMOS) scale.

Task Instructions
In this task, you will hear two samples of speech recordings, one from each system.
The purpose of this test is to evaluate the difference in quality between the two files.
Specifically, you should assess the quality and intelligibility of each file in terms of
its overall sound quality and the amount of mumbling and unclear phrases in the recording.

You should give a score according to the following scale: -3 (System 2 is much worse)
-2 (System 2 is worse)
-1 (System 2 is slightly worse)
0 (No difference)
1 (System 2 is slightly better)
2 (System 2 is better)
3 (System 2 is much better)

Table 9: Similarity mean opinion score (SMOS) Instruction

Introduction
Your task is to evaluate how similar the two speech recordings sound in terms of the speaker's voice.

Task Instructions
In this task you will hear two samples of speech recordings.
The purpose of this test is to evaluate the similarity of the speaker's voice between the two files.
You should focus on the similarity of the speaker,
speaking style, acoustic conditions, background noise, etc.

You should give a score according to the following scale:
5 (Very Similar)
4 (Similar)
3 (Neutral)
2 (Not very similar)
1 (Not similar at all)

H.1 Demographics of Human Evaluators

To assess the quality of synthesized speech, we conducted quantitative evaluation with total of 25 participants. Participants were recruited for individuals physically and cognitively capable of normal activities with ages between 20 and 45 years with high proficiency in English. Recruitment and study procedures adhered to Institutional Review Board guidelines, and all participants provided informed consent. Additionally, all participants were general listeners with no prior expertise in audio or speech synthesis.

H.2 Evaluation Conditions

All participants completed a brief instructive session with an evaluator to familiarize themselves with the evaluation criteria. Evaluation was conducted in a quiet enclosed environment with the same listening device and volume levels, under the instructions of Table 8 and Table 9. Each evaluation took less than 10 minutes.

I Experiment on Unlearning Robustness

While Table 1 shows that TGU has effectively unlearned in overall, we go through extensive experiments to evaluate unlearning robustness. Figure 6 illustrates how TGU unlearned model behaves on remain speakers' speech prompts with various similarity scores to a forget speaker's speech prompt. As unlearning specifically on forget speakers is our objective in speaker identity unlearning, we expect the model to clearly classify forget speakers and remain speakers despite possible resemblances of each other.

For the x-axis, we identified speech prompts in remain set and the highest speaker similarity (SIM) score with any forget speech prompt. Then, the same remain speech prompts were used to generate speech with TGU unlearned model. The y-aixs was then obtained, by comparing the speech prompt with its TGU generated output speech. The results are visualized on 6.

A Pearson correlation analysis was conducted to assess the relationship between the similarity of remain speech prompts to forget speech prompts (x-axis) and the similarity of remain speech prompts to TGU-generated speech output (y-axis). The obtained statistic is 0.1396 while the p-value is 0.0003. This indicates a weak positive correlation with statistical significance, meaning that TGU generated speech is generally independent of the remain samples' similarity to forget speakers. Had the model not been robust and mistreated remain samples as forget speaker samples, there would have been a strong negative correlation.

In Table 10, we further assess the model's robustness by comparing its behavior on remain speakers with similar vocal characteristics to forget speakers. First, we compute the speaker similarity (SIM) between utterances from the two groups. We select utterances whose similarity to any forget-speaker utterance exceeds 0.40 and use these as prompts in our evaluation. Even when a remain speaker's voice closely resembles that of a forget speaker, TGU maintains its performance with the original model. This demonstrates TGU's ability to preserve the identity of remain speakers while effectively neutralizing traces of forget speakers. In constrast, NG, KL and SGU significantly drop SIM-R (drops of 0.226–0.321), suggesting unlearning using these methods may trigger a trade-off that sacrifices remain speaker speech synthesis.

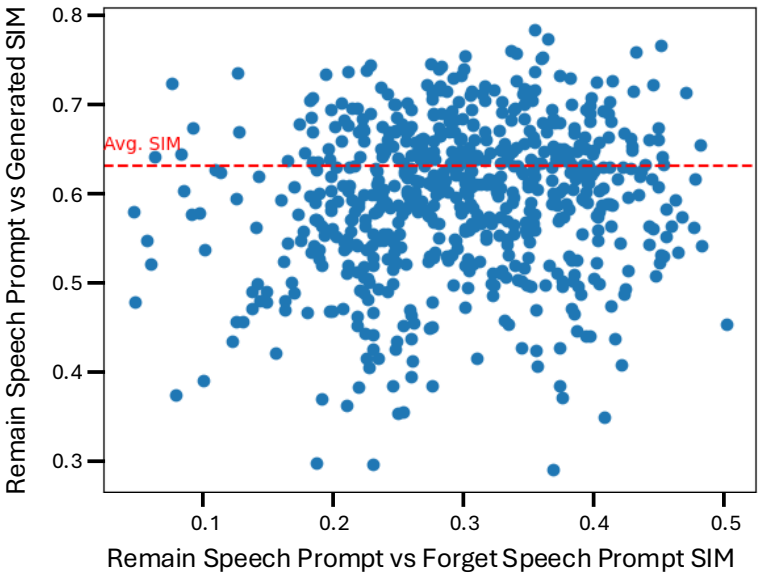


Figure 6: Robustness scatterplot of TGU on remain speakers. The x-axis represents the maximum SIM score between the remain speech prompt and forget speech prompt to depict the level of similarity between a remain speaker and a forget speaker. The y-axis represents the similarity score between the remain speech prompt and its resulting generated output using TGU. The red dashed line indicates average SIM score for all remain speech prompts in the evaluation set.

Table 10: Quantitative results on LibriSpeech test-clean evaluation set (-R) which show high speaker similarity to any forget speaker utterances (exceeding 0.40 in SIM).

Methods	**WER-R** ↓	**SIM-R** ↑
Original	4.96	0.637
NG	6.67	0.393
KL	8.78	0.316
SGU (ours)	5.70	0.411
TGU (ours)	**4.70**	**0.622**
Ground Truth	2.94	-

J Experiment on General Tasks

To provide deeper insights on how TGU unlearning may affect model performances on general tasks where ZS-TTS is used, we experiment the original model and TGU on transient noise removal.

J.1 Transient Noise Removal

ZS-TTS can be applied in tasks where editing is required to remove undesired noise in speech datasets. To prevent having to go through repetitive and inefficient recording to obtain clean speech, ZS-TTS can generate clean audio for the noisy segment. We follow experimental settings of [25] to analyze how TGU unlearned model performs on the task of transient noise removal.

From LibriSpeech test-clean dataset samples of durations 4 to 10 seconds, we construct noise at a -10dB signal-to-noise ratio over half of each sample's duration. Table 11 suggests that TGU provides comparable performances to that of the original model. While seemingly low, diminished model performances on transient noise removal is present relatively to the original model. We suggest that this is a trade-off from successful unlearning. While the model has unlearned to generate voice characteristics of the forget dataset, smaller knowledge-base and implemented randomness could have affected its reconstructing abilities.

Table 11: Transient noise removal results on LibriSpeech test-clean set

Methods	WER↓	SIM↑
Clean speech	4.3	0.689
Noisy speech	47.9	0.213
Original	**2.4**	**0.666**
TGU (ours)	2.5	0.641

Table 12: Diverse speech sampling results on LibriSpeech test-other evaluation set

Methods	WER ↓	FSD ↓
Ground truth	4.5	164.4
Original	8.0	**170.2**
TGU (ours)	**7.9**	177.8

J.2 Diverse Speech Sampling

Being able to generate diverse speech is also an important feature of ZS-TTS models as it ensures realistic and high-quality speech that resembles natural distributions. This is necessary in applications such as speech synthesis or generating training data for speech related tasks (e.g., Automatic Speech Recognition). The diversity of generated speech samples is measured with Fréchet Speech Distance (FSD) as suggested in [25]. From generated speech samples, we extracted self-supervised features using 6th layer representation of wav2vec 2.0 [1]. The features were reduced to 128 dimensions with principle component analysis and used to calculate the similarity of distributions with real speech. High FSD indicates lower quality and minimal diversity, while low FSD refers to high quality and more diversity. For this experiment, α is set to 0 to ensure more diversity. Ground truth FSD is obtained by partitioning the LibriSpeech test-other set into half while ensuring equal distribution of data per speaker across both subsets

Experimental results in Table 12 show that FSD increases in TGU unlearned model. Because this task does not require input audio prompts, diverse speech sampling relies relatively heavier on datasets used to train the model. Implementing machine unlearning and thus inducing forgetting of specific speakers causes a trade-off in model's diversity. Meanwhile, it is noticeable that TGU achieves a lower WER in this case. We can infer that TGU obtains robustness in relatively noisy dataset comparable to the Original model.

K Recovery Experiment

Table 13 illustrates an experimental result on whether an unlearned model is recoverable to its original state. Aligning with our motivation to make ZS-TTS models safe, we presume a scenario of a privacy attacker who attempts to retrieve the original model parameters. We train the TGU unlearned checkpoints on all 10 of forget speaker's dataset to recover the original model. We also presume a practical scenario and attempt to recover the model performance using average of 1 minute for each speaker.

When given audio duration of 15 minutes for the forget speakers, the model fails to generalize over other speakers, hence, failing to mimic voices other than the forget speaker's. Additionally, the recovered model is more likely to generate wrong speech content as shown with higher WER in both remain set and the forget set. This process resembles fine-tuning a Text-to-Speech model for specific speakers rather than true recovery. Consequently, the original ZS-TTS model cannot be restored, and the attacker is essentially leveraging transfer learning to create a forget speaker-specific TTS model. However, with enough training data, the attacker could achieve similar results using any other non-zero-shot TTS model. We also consider a scenario where an attacker has access to only 1 minute of the forget speaker's voice sample. In this case, the model parameters also remain unrecoverable. The model also fails to generate forget speaker's voice. The model loses its zero-shot abilities hence the performance at early steps. Therefore, in practical scenarios where an attacker may attempt to train the model to clone an individual's voice with short sample of speech (e.g., voice phishing), it would not be feasible to recover the model or successfully generate the forget speaker's voice.

Table 13: Quantitative results for recovery experiments on unlearned models. WER and SIM evaluation follows the procedures of Table1.

Methods	**Recover Steps**	**Audio per Spk**	**WER-R** $\downarrow$	**SIM-R** $\uparrow$	**WER-F** $\downarrow$	**SIM-F** $\uparrow$
Original	-	15 min	2.1	0.649	2.1	0.708
TGU	-	15 min	2.5	0.631	2.4	0.169
TGU	36.25K	15 min	4.23	0.303	2.5	0.735
TGU	14.5K	1 min	4.61	0.226	2.8	0.162

Table 14: Quantitative results for reproducibility experiments using LibriTTS as pre-train dataset.

Methods	**Unlearn Steps**	**WER-R** $\downarrow$	**SIM-R** $\uparrow$	**WER-F** $\downarrow$	**SIM-F** $\uparrow$
Original	-	3.2	0.610	6.3	0.503
TGU	10K	3.3	0.548	6.4	0.184

L Reproducibility Experiment

Table 14 illustrates the reproducibility of our experiment of Table 1 using a different dataset, LibriTTS [55]. We pre-trained the VoiceBox model on LibriTTS for 500K steps and 10 speakers were randomly selected as forget set. When performing unlearning for only 10K steps (7% of pre-training steps), results of TGU illustrate effective unlearning while maintaining content accuracy.

M Limitations and Discussions

M.1 Limitations

While we were able to effectively apply machine unlearning to ZS-TTS to suppress replication of specific identities, we also acknowledge potential trade-offs or untackled challenges.

In J, we have investigated the difference in performance of the TGU model and original model on downstream tasks. On downstream usage of unlearned ZS-TTS model, machine unlearning, in its nature, removes knowledge or datasets from the model's parameters. Therefore, the impact of unlearning process could pose unexpected results when the model is used for various purposes. In K, we analyze the possibility of regaining unlearned model parameters to its original state - an attack scenario of malicious user. Without usage of sufficient remain set, the original model weights of ZS-TTS system cannot be recovered. However, unlearning does not safeguard the architecture itself from being trained to replicate a specific forget speaker after the training process. This challenge should be tackled in future works, where adaptation of methods to hinder further finetuning on unwanted speaker identities can be applied for voice privacy.

In scope of claims, our method requires access to the original training set and becomes more computationally demanding as the number of forget speakers grows. Without access to the original training set (or other sufficient remain dataset), the unlearning performance may not be consistent and could be prone to catastrophic forgetting. This reflects a broader challenge in machine unlearning, where efficiency and practicality remain open questions. As the first work to address unlearning in ZS-TTS, our focus is on establishing a strong foundation. We anticipate future research will extend this direction by developing lighter approaches and exploring scenarios with more limited data access.

This paper provides results in different settings, using a VoiceBox model pretrained on LibriHeavy in 1 and on LibriTTS in L. We selected VoiceBox for its state-of-the-art performance at the time of our study. While our evaluations focus on this architecture, the approach is expected to extend naturally to other flow-matching models with analogous structures, such as F5TTS, which we anticipate will exhibit similar behavior. To further reinforce the generality of our findings, additional experiments could be made on more ZS-TTS systems.

M.2 Discussions

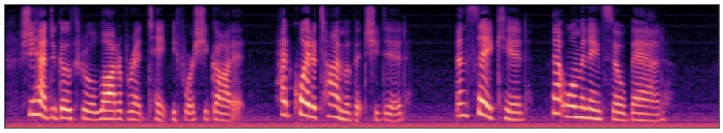
Although our work addresses the need for individuals to opt out of voice replication, determining how to handle similar voices raises complex questions. In striving to protect the privacy of a single individual, one could unintentionally restrict beneficial TTS capabilities to others whose voices resemble the forget set. Balancing personal privacy rights and broader technological benefits is at the heart of this tension. Also, techniques for

ensuring speaker identity unlearning must be verifiable and transparent. Providing evidence that the model no longer replicates a forgotten identity requires both quantitative evaluation and subjective analysis. In light of the current situation—where many models remain closed due to concerns about misuse—we believe our work marks a new chapter in safeguarding individuals, paving the way for broader availability in the future. We aim to foster a deeper ethical discourse and encourage further research on responsibly handling ZS-TTS.

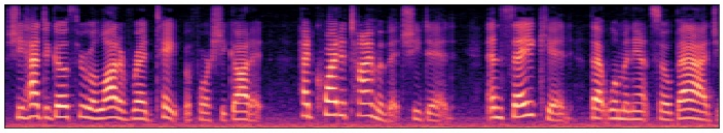
Overall, this work establishes the first foundation for applying machine unlearning to ZS-TTS. While certain aspects such as efficiency, robustness across diverse conditions, and fairness in privacy protection remain open challenges, we view these as important directions for future research. By addressing them, subsequent studies can build upon our framework to develop more practical, resilient, and equitable unlearning methods.

N Inference Samples

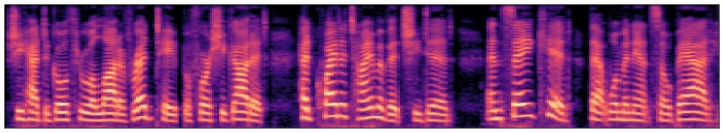
Figures 7 and 8 show the Mel-spectrograms for the ground truth, original VoiceBox, SGU, and TGU inference results on forget speaker samples. These figures represent samples from speakers *789* and *6821*, respectively. The ground truth Mel-spectrogram corresponds to the audio where the same speaker as the prompt reads the same transcription.



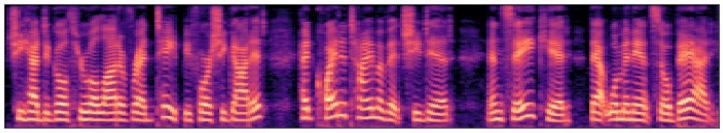
(a) Ground Truth



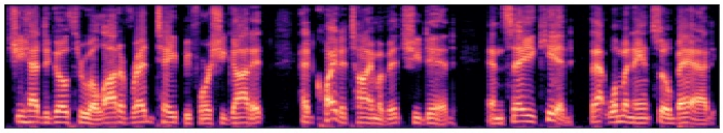
(b) Original



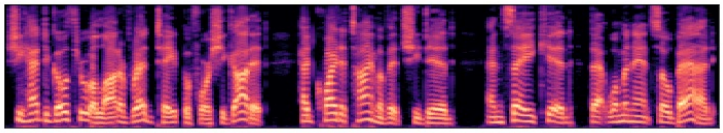
(c) SGU Sample 1



(d) SGU Sample 2

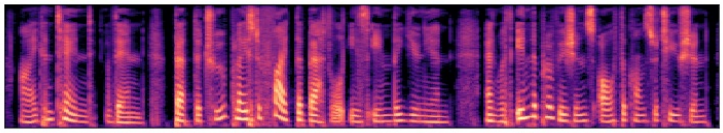


(e) TGU Sample 1

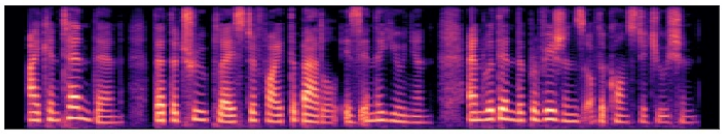


(f) TGU Sample 2

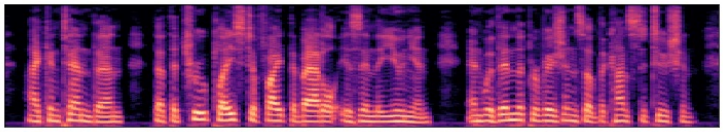
Figure 7: Mel-Spectrogram Comparisons: GT, Original, SGU Samples, and TGU Samples for the forget speaker *789*



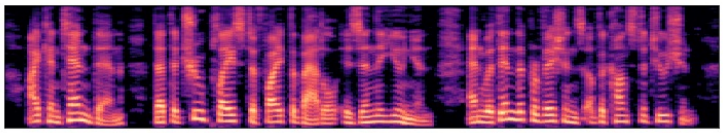
(a) Ground Truth



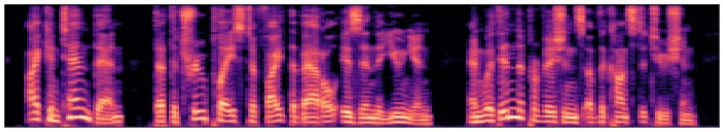
(b) Original



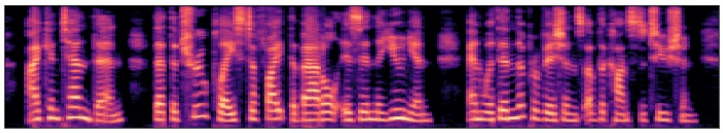
(c) SGU Sample 1



(d) SGU Sample 2



(e) TGU Sample 1



(f) TGU Sample 2

Figure 8: Mel-Spectrogram Comparisons: GT, Original, SGU Samples, and TGU Samples for the forget speaker *6821*

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and introduction both include the claims made in this paper - specifically, suggested method of Guided Unlearning and novel metric spk-ZRF. These methods are followed by experimental results that prove our proposed methods' effectiveness in unlearning speaker identities in ZS-TTS system.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix M, we discuss the limitations of work in terms of assumptions, scope of claims, and factors that may influence performance followed by ethical considerations of the work. Additionally, we provide sufficient discussions in robustness in Appendix I, scalability in terms of forget set size in Table 2, computational efficiency in Appendix E. We believe these extensive analysis and transparent reflections of work further strengthen not only this paper, but future works in speaker identity unlearning.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not contain any theoretical claims.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper clearly states with pretrained model (VoiceBox [25]) was used, and usage of publicly available and cited datasets (LibriHeavy [23], LibriTTS [55], LibriSpeech [22]. We provide implementation details of all baselines and our methods in full detail with hyperparameters in Appendix B. We also report the model and inference configurations, data preprocessing, remain/forget split sizes, selection of forget speakers, evaluation settings, and pre-training settings in Appendix A. The proposed evaluation pipeline and unlearning pipelines is made publicly available as well. Yet, we do not redistribute pre-trained VoiceBox model weights nor the VoiceBox pre-training code in compliance the Ethics Statement of original authors' measures due to risks of misuse[3]. Nevertheless, researchers can reproduce the experiments following publicly available procedures to reproduce VoiceBox pre-training along with our provided pre-training procedures.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

[3] https://voicebox.metademolab.com/

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code for data access and preparation, evaluation pipeline and training pipeline of unlearning methods. As mentioned in the previous checklist, we omit pre-training code and weights for reproducing VoiceBox hence privacy. Additionally, we provide generated results using our methods in the project page.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify full details on training and test settings in terms of data splits, optimizers, hyperparameters in not only the Section 5.1, but also in Appendix A, Appendix B, and ablations in Appendix G.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide statistical significance of experiments in Appendix F for main claims of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the type of compute workers with relevant memory for the main experiment and estimate of compute in Appendix D

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: For human assessment, we follow the guidelines of IRB of the institution and specify the procedures as mentioned in Appendix H.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the impacts and ethical considerations of this paper that may result in unwanted results; where remain speakers of highly similar identities to that of forget speakers may be omitted from accessibility to ZS-TTS system usage in Appendix M. We make effort to cover this part by evaluation on similar speakers in Appendix I.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the

other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We release all code but omit the weights and training code for the VoiceBox model which has a high risk of misuse (speech generation without consent).

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the creators or owners of all datasets, code, and models used in this paper in both the main content and specify in Appendix A.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: While this paper does not release new dataset or models, we release the pipelines of proposed method and code for evaluation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We provide the full text of instructions and research procedures with human subjects, along with details about compensation in Appendix H.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: We obtained IRB approval for experiments that involve human subjects in Table 4 and mention this in Appendix H.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We do not use LLMs in any way impacting important, original, or non-standard component of the core methods in this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.