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

Anonymous Authors[1]

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. The demo is available at `https://speechunlearn.github.io/`.

1. Introduction

Significant advancements in Zero-Shot Text-to-Speech (ZS-TTS) (Le et al., 2024; Casanova et al., 2022; Ju et al., 2024; Wang et al., 2025) enable models to synthesize speech accurately using minimal speaker input. Methods like VALL-E (Wang et al., 2025) utilize discrete speech tokens, while VoiceBox (Le et al., 2024) employs masked prediction for speech synthesis and audio infilling. Given that a person's voice is a key biometric characteristic used for identification (Nautsch et al., 2019a;b), 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.

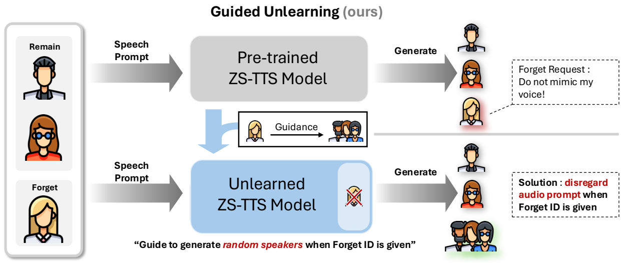

Guided Unlearning (ours)
Remain
Forget
Speech Prompt
Pre-trained ZS-TTS Model
Generate
Forget Request : Do not mimic my voice!
Guidance
Speech Prompt
Unlearned ZS-TTS Model
Generate
Solution : disregard audio prompt when Forget ID is given
"Guide to generate random speakers when Forget ID is given"

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.

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 (Panariello et al., 2024; Tomashenko et al., 2024), and thus MU has gained traction across various fields of generative AI. Despite growing privacy concerns in speech-related tasks (Tomashenko et al., 2022; Yoo et al., 2020), 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 approaches : computationally efficient Sample-Guided Unlearning (SGU) and advanced Teacher-Guided Unlearning (TGU). Guided Unlearning incorporates randomness into voice styles when encountering forgotten speakers, ensuring voice neutralization while preserving synthesis quality for other speakers. We also propose a novel metric, speaker-Zero Retrain Forgetting (spk-ZRF), to measure randomness in voice generation for forget prompts, enhancing privacy evaluation. TGU notably achieves a 2.95 % increase in speaker identity randomness compared to the baseline.

The main contributions are:

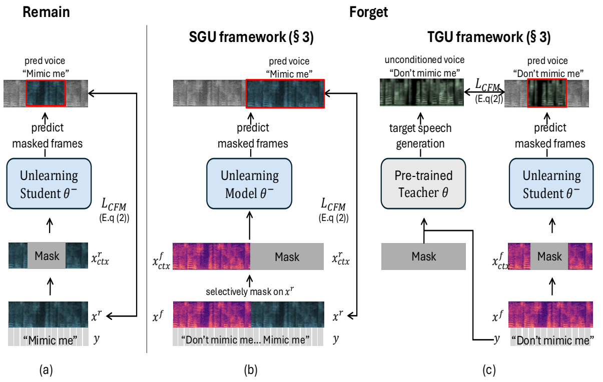


Figure 2. The training procedure for the forget set in (b) the SGU framework and (c) the proposed TGU framework, along with (a) the training procedure for the remain set in both SGU and TGU.

- Introducing the first speaker identity unlearning framework specifically for ZS-TTS.
- Proposing SGU and TGU frameworks that effectively reduce voice replication capabilities for forgotten speakers.
- Introducing spk-ZRF, a metric evaluating the effectiveness of unlearning by measuring randomness in generated speaker identities.

2. 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. A typical ZS-TTS model θ takes as input a pair (x^s, y), where x^s is a speech prompt from speaker $s \in S$ and y is the corresponding text. The model generates synthesized speech $\hat{x}_y^{spk=s}$ that delivers y in the voice style of speaker s:

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

Given a pre-trained ZS-TTS model θ trained on data D^S, we aim to construct an unlearned model θ^- that satisfies the following two objectives:

Retain Objective. For any speaker $r \in R$, the model should behave as before, generating the given text y in the voice of r:

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

Forget Objective. For any speaker $f \in F$, the model should still generate speech for the given text y, but using a voice that is different from that of speaker f:

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

That is, the model should no longer replicate the voice characteristics of any speaker in the forget set F. Moreover, the generated speech should not exhibit a fixed or consistent style that could indirectly reveal the forgotten identity. This formulation ensures both the preservation of TTS capabilities and the protection of speaker privacy.

3. Method

To explore unlearning in ZS-TTS, we adopt VoiceBox (Le et al., 2024), a state-of-the-art text-guided non-autoregressive (NAR) model for multilingual speech synthesis and editing. VoiceBox leverages Conditional Flow Matching (CFM) to transform a prior distribution p_0 (e.g., Gaussian) into the target speech distribution p_1 by learning a conditional vector field $v_t(w, y, x_{ctx}; \theta)$, where $w = (1-(1-\sigma_{\min})t)x_0 + tx$ and $x_{ctx} = (1-m) \odot x$ is the masked input. The model is trained by minimizing the discrepancy between the predicted vector field and a guiding field $u_t(x|x_1)$ through the following loss:

$$\begin{aligned} L_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)} \Big[\| m \odot u_t(x|x_1) \\ - v_t(w, y, x_{ctx}; \theta)\|^2 \Big], \end{aligned} \tag{4}$$

where $p_t(x|x_1) = \mathcal{N}(x|tx_1, \sigma_t^2 I)$ is a Gaussian path with time-dependent variance $\sigma_t = 1-(1-\sigma_{\min})t$. This approach allows VoiceBox to model speaker style in a label-free manner.

To guide the model toward generating text y in a random, non-identifiable voice style, one would ideally use paired data $(x^{spk\neq f}, y)$, where $x^{spk\neq f}$ is an utterance of y in a voice different from the forget speaker f. However, such aligned pairs are not naturally available.

As a workaround, we propose **Sample-Guided Unlearning (SGU)**: for a given text y, we concatenate (x^r, y) from a remain speaker and (x^f, y^f) from a forget speaker to form a pseudo-sample. We then mask x^r entirely and use it as the target for infilling (Figure 2-(b)). However, unlike the original VoiceBox setting where both preceding and succeeding contexts are available, SGU provides only one-sided context, limiting effective infilling. Furthermore, if masking is applied mid-sequence, speaker mismatches in prosody or rhythm may lead to unnatural synthesis and degraded quality.

To overcome the limitations of SGU, we propose **Teacher-Guided Unlearning (TGU)**, which leverages the pre-trained model θ to generate text-speech aligned targets for unlearning. Specifically, we utilize the property that $\theta(y)$—when conditioned only on text—generates speech in varying voice styles depending on the Gaussian initialization x_0, ensuring non-identifiability:

$$\theta^-(x^f, y) \approx \theta(y). \tag{5}$$

As illustrated in Figure 2-(c), given (x^f, y), we use $\bar{x} = \theta(y)$ as the target for model θ^-, which is initialized from θ. The forget loss is defined as:

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

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

To preserve performance on non-forget speakers, we apply the original CFM loss on the remain set D^r (Figure 2-(a)):

$$L_{\text{CFM-remain}}(\theta^-) = \mathbb{E}_{t,q(x_1),p_t(x^r|x_1)}\Big[\|m \odot u_t(x|x^r_1) - v_t(w^r, y, x^r_{ctx}; \theta^-)\|^2\Big], \tag{7}$$

with w^r defined identically to w.

The final objective combines both losses:

$$L_{\text{total}} = \lambda L_{\text{CFM-remain}} + (1 - \lambda) L_{\text{CFM-forget}}, \tag{8}$$

where λ balances the trade-off and is set to 0.2.

3.1. Proposed Metric: spk-ZRF

Conventional machine unlearning (MU) metrics—such as completeness (Wang et al., 2024), JS-divergence, or activation distances—primarily assess performance gaps between forget and remain sets. However, such comparisons can be misleading, as consistent patterns in the forget set may still allow reverse-engineering of the forgotten data.

To address this, we propose a novel metric named **speaker-Zero Retrain Forgetting (spk-ZRF)**, which quantifies the randomness of speaker identity generation in ZS-TTS, independent of speech content quality. Inspired by Zero Retrain Forgetting (ZRF) (Chundawat et al., 2023), spk-ZRF adapts the idea for voice identity by comparing the model's outputs with and without speaker prompts.

Given a test set $D^S = \{(x^s_{y_i}, y_i)\}_{i=1}^n$, we compute speaker embeddings of $\theta^-(x^s_i, y_i)$ and $\theta(y_i)$ using a speaker verification model, then convert them into probability distributions via softmax. The Jensen-Shannon divergence (JSD) between the two embeddings is:

$$\begin{aligned} \text{JSD}_i = {} & 0.5\, D_{\text{KL}}(\text{Softmax}(_{\theta(x^s_i, y_i)}) \parallel M_i) \\ & + 0.5\, D_{\text{KL}}(\text{Softmax}(_{\theta(y_i)}) \parallel M_i), \end{aligned} \tag{9}$$

where M_i is the average of both distributions:

$$M_i = \frac{1}{2}\left(\text{Softmax}(_{\theta(x^s_i, y_i)}) + \text{Softmax}(_{\theta(y_i)})\right). \tag{10}$$

The overall spk-ZRF score is given by:

$$\text{spk-ZRF} = 1 - \frac{1}{n}\sum_{i=1}^{n} \text{JSD}_i. \tag{11}$$

A spk-ZRF closer to 1 indicates high randomness in generated speaker identity, implying successful unlearning. In contrast, a lower score reflects persistent identity patterns, suggesting incomplete forgetting.

4. Experimental Setup

Baseline Methods. We evaluate four unlearning baselines applied to VoiceBox (Le et al., 2024). (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 (Warnecke et al., 2021). (3) **Negative Gradient (NG)** performs gradient ascent on the forget set D^F (Thudi et al., 2022; Fan et al., 2024). (4) **Selective KL Divergence (KL)** maximizes KL divergence for forget samples while minimizing it for remain samples using a teacher model (Li et al., 2024; Chen & Yang, 2023).

Dataset and Model Configuration. We adopt a realistic setting by unlearning multiple speakers simultaneously, unlike prior works that focus on a single identity (Gandikota et al., 2023; Seo et al., 2024). Experiments are conducted on LibriHeavy (Kang et al., 2024), a 50K-hour English corpus with transcriptions. In Table 1, we unlearn 10 randomly chosen speakers (20 minutes each). Each speaker has 5 minutes for evaluation and the rest for training. VoiceBox is trained on LibriHeavy with consistent configurations. For generalization to unseen speakers, we use LibriSpeech test-clean (Panayotov et al., 2015).

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 (Hsu et al., 2021) 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 (Chen et al., 2022). For qualitative evaluation, we use Comparative MOS (CMOS) for audio quality and Similarity MOS (SMOS) for voice similarity. Additional training and inference details are provided in Appendix A.

4.1. 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 2), 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 gener-

Table 1. Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set (-F). $^\diamond$ 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. The result of ANOVA test on JSD, which was averaged to calculate spk-ZRF, indicated significant differences in spk-ZRF across remain set ($F(4, 768) = 116.31, p < 0.0001$) and forget set ($F(4, 1188) = 807.97, p < 0.0001$) among models.

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
FT	**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	-	-	-

ating noise instead of distinct voices due to entanglement between style and content.

Among all methods, TGU best 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 values.

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.

Human Subjective Evaluation. Table 2 reports qualitative results via CMOS and SMOS. Compared to SGU, TGU achieves closer CMOS scores to the original model, confirming its ability to maintain natural speech quality. For SMOS, TGU outperforms SGU in preserving voice styles for remain speakers and generates more distinct outputs for forget prompts, effectively preventing voice replication.

Table 2. Human assessment on Librispeech test-clean evaluation set (R) and the forget evaluation set (F).

Methods	CMOS		SMOS	
	R ↑	F ↑	R ↑	F ↓
OG	0.00 ± 0.0	0.00 ± 0.0	4.47 ± 0.4	4.44 ± 0.4
SGU	-0.15 ± 0.3	-0.53 ± 0.3	3.12 ± 0.8	1.45 ± 0.3
TGU	**-0.02 ± 0.2**	**-0.45 ± 0.2**	**4.67 ± 0.3**	**1.28 ± 0.2**
GT	1.00 ± 0.26	0.22 ± 0.29	3.70 ± 0.7	3.89 ± 0.7

These findings support that TGU restricts imitation of forget speakers while maintaining the ZS-TTS model's usability. See Appendix F for details.

5. 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.

References

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.

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.

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

Chen, R. T. Q. torchdiffeq, 2018.

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.

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. URL https://doi.org/10.1609/aaai.v37i6.25879.

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.

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.

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

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.

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.

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.

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.

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.

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

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.

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*, 2019a.

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, 2019b.

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.

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.

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.

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.

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.

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.

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.

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

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.

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

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.

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.

A. Experiment Settings

A.1. Dataset Details

For the training set, we utilized the LibriHeavy dataset ((Kang et al., 2024)), 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 ((Panayotov et al., 2015)). It is important to note that there is no overlap between the speakers in the LibriSpeech test-clean set and those in LibriHeavy ((Kang et al., 2024)). Following the experimental setup outlined in the original VoiceBox paper ((Le et al., 2024; Wang et al., 2025)), 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) (McAuliffe et al., 2017) 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 (Le et al., 2024), using the same configuration. The audio feature generator is based on a vanilla Transformer (Vaswani, 2017), enhanced with U-Net style residual connections, convolutional positional embeddings (Baevski et al., 2020), and AliBi positional encoding (Press et al., 2022). 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 (Le et al., 2024). 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 (Kong et al., 2020), trained on the LibriHeavy (Kang et al., 2024) English speech dataset, is employed to convert the spectrogram into a time-domain waveform.

A.5. Pre-training

Following (Le et al., 2024), 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, (Ho & Salimans, 2022; Le et al., 2024)) was applied as follows:

$$\hat{v}_t(w, x, y; \theta) = (1 + \alpha) \cdot v_t(w, x_{ctx}, y; \theta) - \alpha \cdot v_t(w; \theta) \tag{12}$$

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

We utilized the `torchdiffeq` package (Chen, 2018), 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.

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 6. 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 **??**, 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 6. 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 (Thudi et al., 2022). 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 (Li et al., 2024; Chen & Yang, 2023). 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 :

$$L_{\text{KL}} = \lambda(\theta(x^r, y^r) \| \theta^-(x^r, y^r)) - (1-\lambda)(\theta(x^f, y^f) \| \theta^-(x^f, y^f)) \tag{13}$$

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

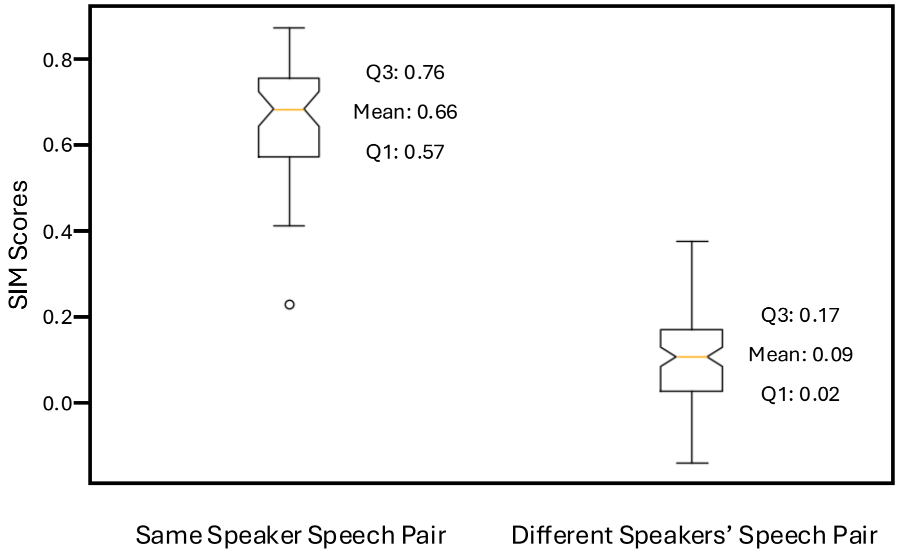
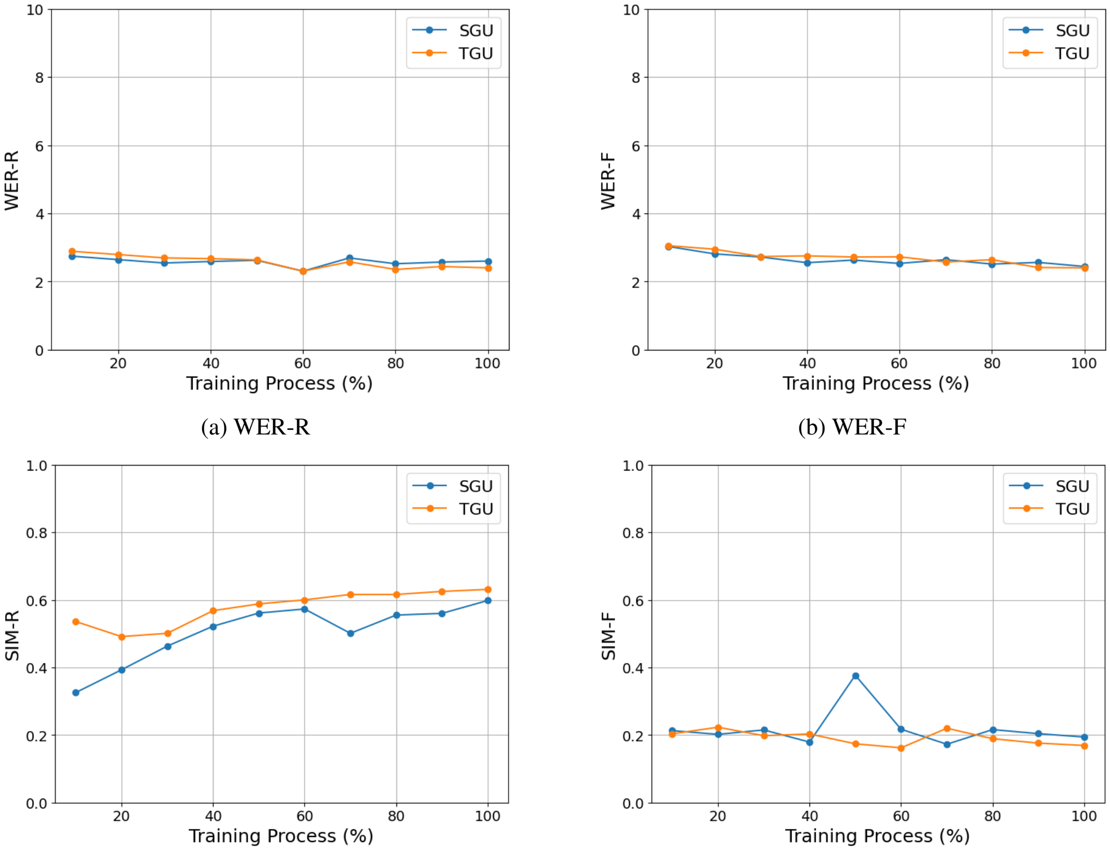


Figure 3. Boxplot of speaker similarity on same speaker's and different speakers' audio. Each are evaluated with 100 pairs of random speech audio in LibriSpeech test-clean subset.

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 3, 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 6. 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. Quantitative results over the training process



(a) WER-R

(b) WER-F

(c) SIM-R

(d) SIM-F

Figure 4. 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 4 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.

E. Impact of α

Table 3. 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 3 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.

F. Qualitative Instruction

Table 4 and Table 5 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 4. 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 5. 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)

F.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.

F.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 4 and Table 5. Each evaluation took less than 10 minutes.

G. Visualization

Figure 5 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. 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 oveall.

H. Scalability

Table 6 shows that both SGU and TGU successfully unlearn the target speakers in both scenarios while preserving intelligibility on the remain set (R). Notably, even when scaling from 1 to 10 speakers, both methods continue to yield solid results, with TGU displaying almost no performance degradation. In contrast, SGU sees 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 irrespective of the number of forgotten speakers.

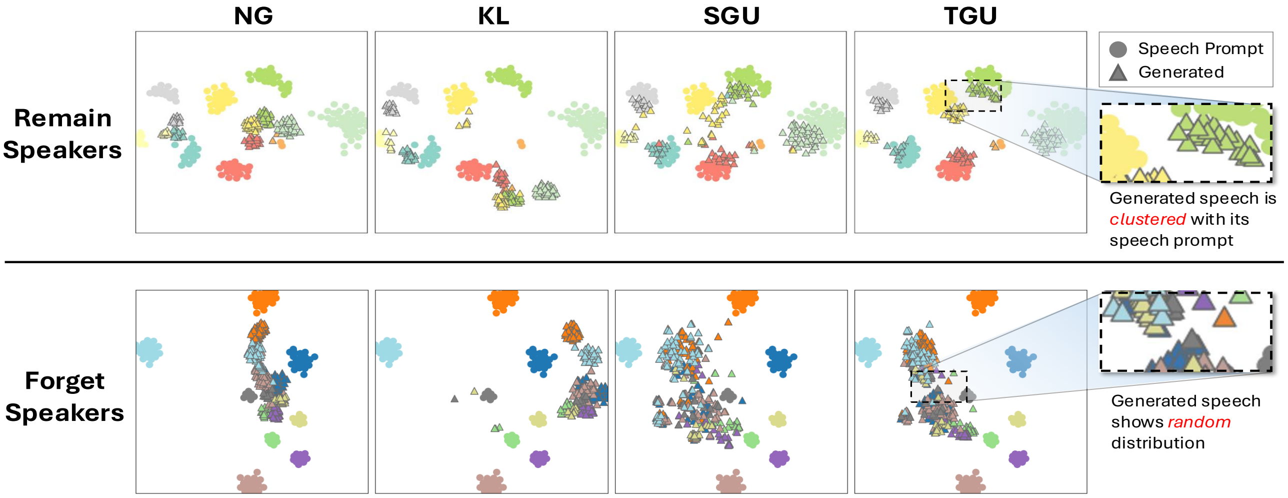


Figure 5. t-SNE analysis comparing NG, KL, SGU, and TGU for remain and forget sets. Samples from the same speaker are represented with the same color, where circles indicate actual speaker embeddings and triangles represent the embeddings of the model-generated speech. Ideal unlearned model should generate 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 6. 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.

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

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.

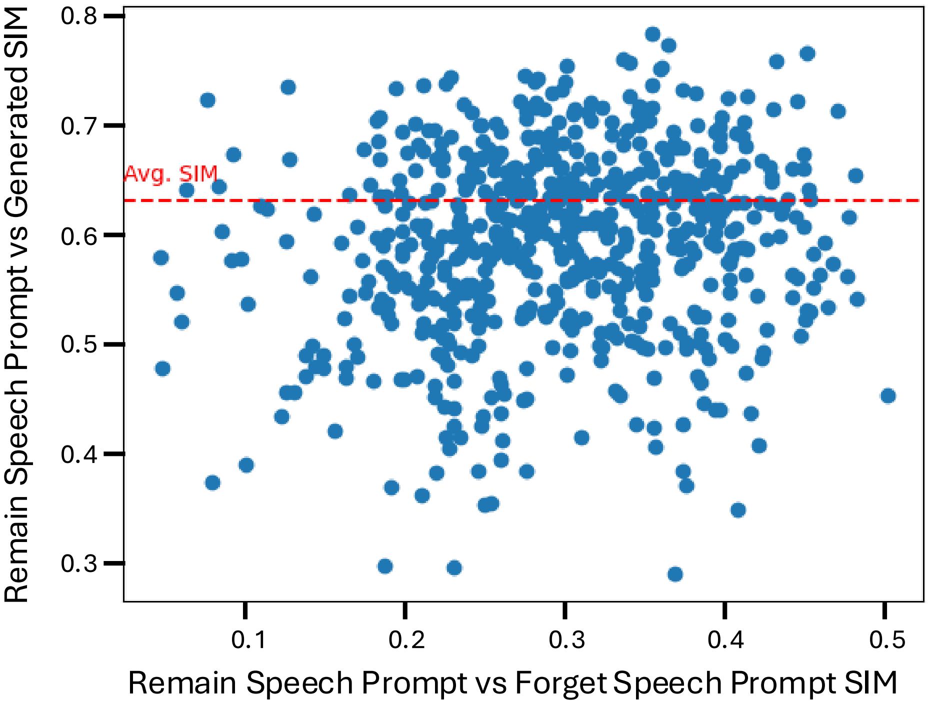


Figure 6. Robustness scatterplot of TGU on remain speakers. The x-axis represents the maximum SIM score between the remain speech prompt and forget speech prompt to depict the level of similarity between a remain speaker and a forget speaker. The y-axis represents the similarity score between the remain 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 7. 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 8. 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. 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 (Le et al., 2024) 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 7 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.

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 (Le et al., 2024). From generated speech samples, we extracted self-supervised features using 6th layer representation of wav2vec 2.0 (Baevski et al., 2020). 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 8 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.

Table 9. Recovery experiment on TGU unlearned model

Methods	Recover Steps	Audio per Spk	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↑
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 10. Quantitative results for recovery experiments on unlearned models. WER and SIM evaluation follows the procedures of Table1.

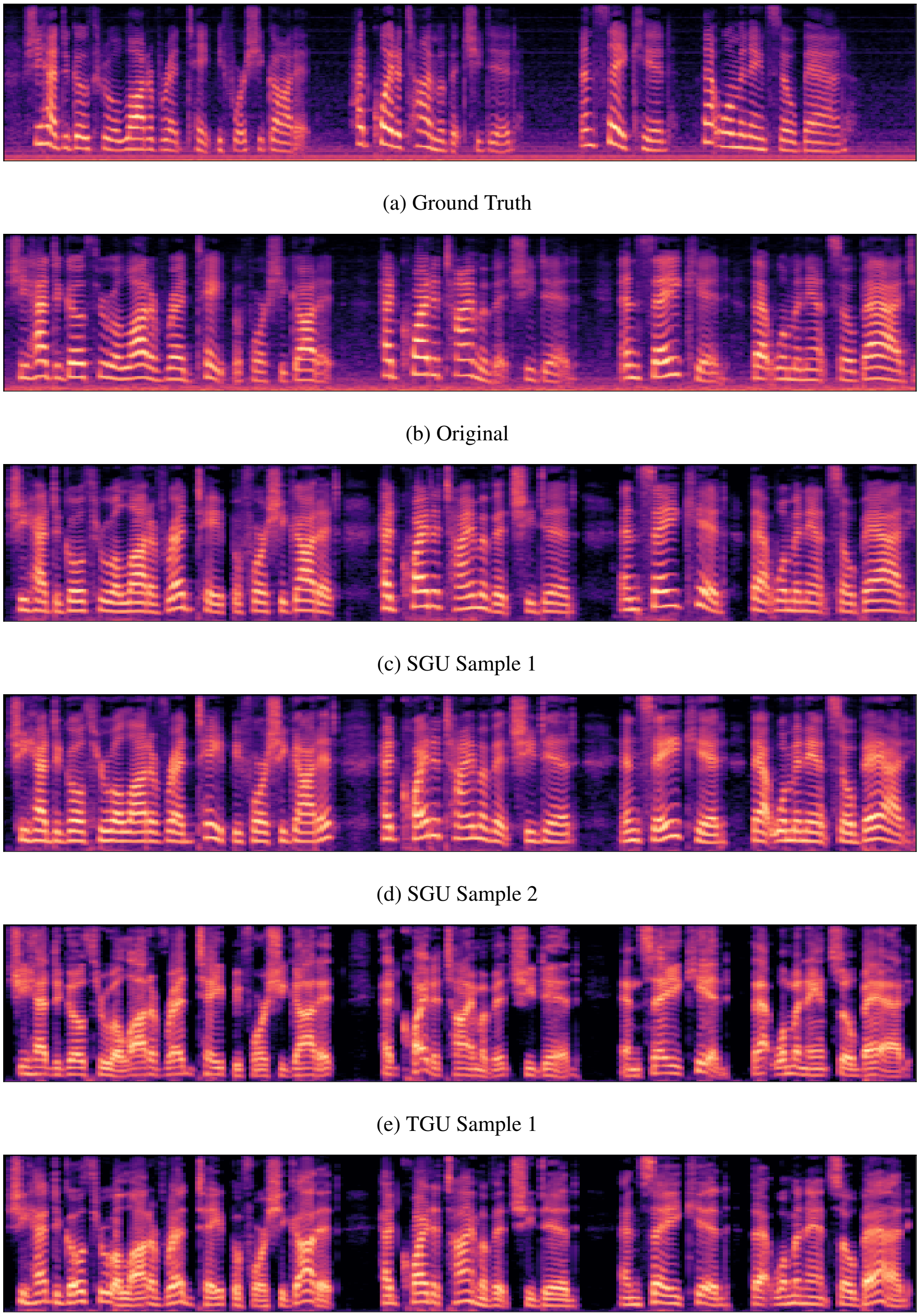
K. Recovery Experiment

Table 10 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, provided sufficient training data for the forget speaker is available. 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. With shorter audio duration, 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.

L. 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

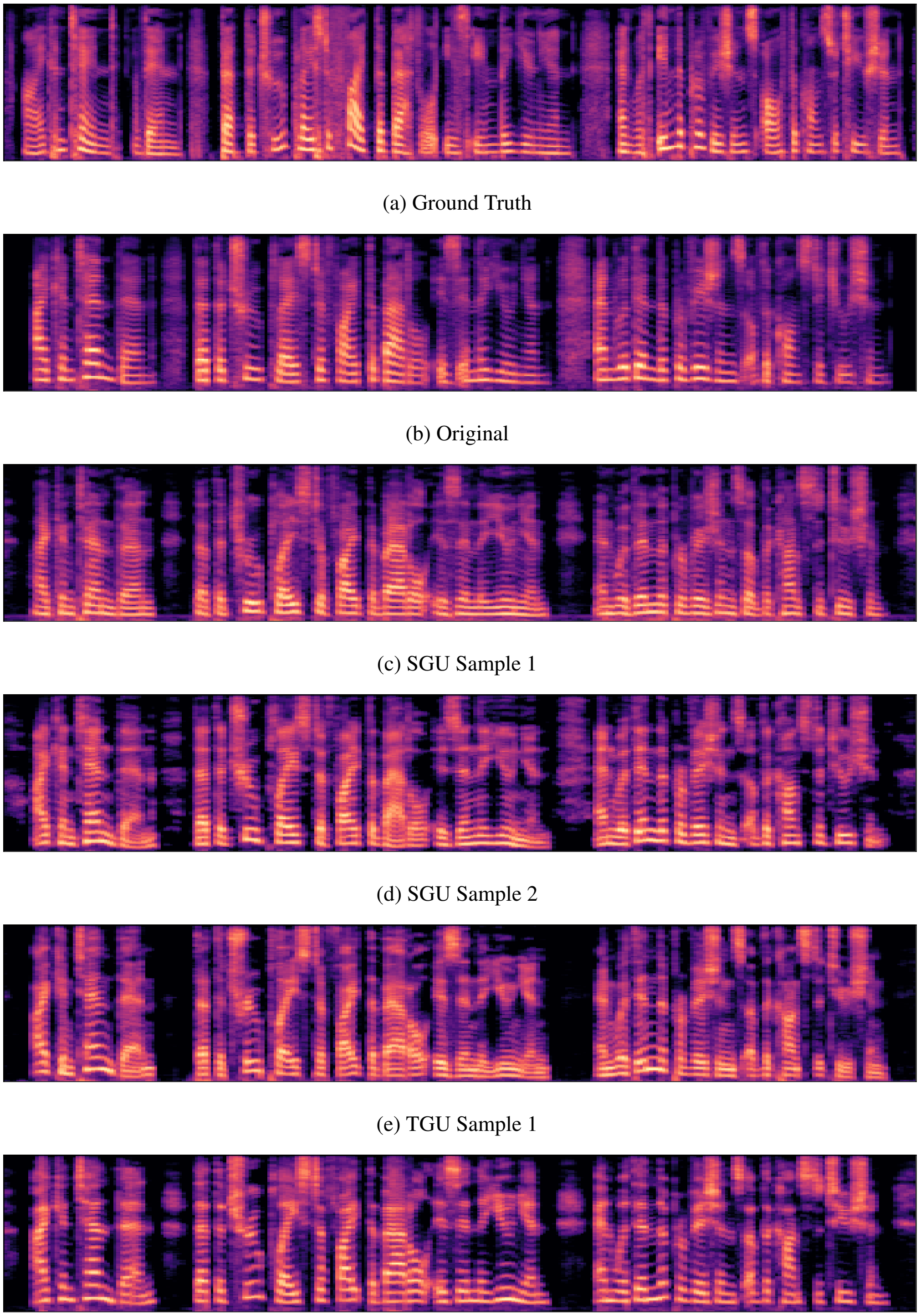
(c) SGU Sample 1

(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*