ERASING YOUR VOICE BEFORE IT'S HEARD: TRAINING-FREE SPEAKER UNLEARNING FOR ZERO-SHOT TEXT-TO-SPEECH

Myungjin Lee[*] *Eunji Shin*[*] *Jiyoung Lee*[†]

Ewha Womans University

ABSTRACT

Modern zero-shot text-to-speech (TTS) models offer unprecedented expressivity but also pose serious crime risks, as they can synthesize voices of individuals who never consented. Existing unlearning approaches, reliant on retraining, are costly and limited to speakers seen in the training set. We present **TruS**, a training-free speaker unlearning framework that shifts the paradigm from data deletion to inference-time control. **TruS** steers identity-specific hidden activations to suppress target speakers while preserving other attributes (*e.g.*, prosody and emotion). Experiments on F5-TTS show that TruS effectively forgets both seen and unseen speakers without retraining, establishing a scalable safeguard for speech synthesis. The demo is available on `http://mmai.ewha.ac.kr/trus`.

***Index Terms*—** Speaker unlearning, text-to-speech (TTS), steering activations

1. INTRODUCTION

The advancement of zero-shot text-to-speech (TTS) recently reached a level of expressivity and naturalness that makes it attractive for a wide range of applications [1, 2]. Meanwhile, their ability to generalize across speakers [3, 4, 5] introduces a profound risk: such models may generate the voices of real individuals who never consented to their use [6, 7]. Prior attempts to mitigate this issue, such as watermarking [8, 9], can only verify or trace synthetic audio after it has been produced. A more extreme solution would be to prohibit voice generation altogether. Yet this would undermine the utility of TTS technology, which is critical for accessibility and HCI.

In this context, machine unlearning has recently emerged to remove the influence of specific training data from learned models. However, most existing methods, including gradient-based approaches [10, 11, 12] or knowledge distillation-based [13, 14], have primarily been developed in the field of image or text generation. These approaches often involve full or partial retraining of the model, leading to high computational cost, unstable convergence, and unintended forgetting. In TTS, only a few attempts have been made [15]: Sample-guided unlearning (SGU) and teacher-guided unlearning (TGU). While TGU adapts distillation to forget training-set speaker identities, it still requires substantial retraining whenever new unlearning requests arise, making it impractical for scalable deployment. However, this approach still incurs considerable training costs and relies heavily on the quality of the teacher model.

In contrast to prior efforts, we introduce the concept of opt-out unlearning, which allows individuals to explicitly request that their voices not be synthesized. This reframes unlearning as a user-driven safeguard: the system is able to suppress the generation of unseen voices when requested. We therefore distinguish between the *forget* set, consisting of training-set speakers marked for removal, and the *forbidden* set, which represents unseen speakers who have opted out, as depicted in Fig. 1.

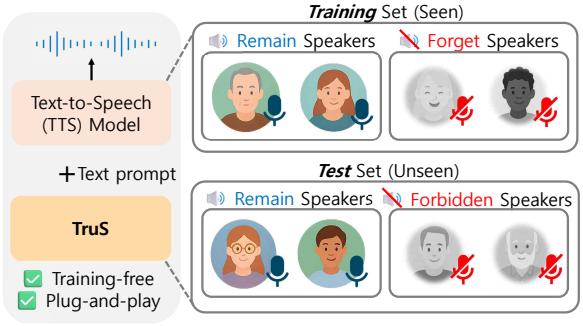


Fig. 1: Illustration of training-free speaker unlearning.

To this end, we propose a simple yet novel training-free steering mechanism for speaker unlearning in TTS, shortly **TruS**. Previous unlearning techniques have learned models to remove trained identities by collapsing into a nearly identical persona [16]. Such over-suppression undermines the practical utility of the system. By contrast, our goal is to prevent the synthesis of specific speaker identities while preserving essential paralinguistic attributes. Recently, EmoSteer [17] has attempted to modulate emotional prosody by heuristically selecting top-k activation channels, but applying the same fixed rule across inputs removes any dynamic adaptability. On the other hand, our **TruS** first generates an identity prototypical embedding with intermediate features of TTS models. We statistically analyze the similarity according to identity in each layer to select steering blocks. At inference, **TruS** dynamically guides hidden representations with such a prototypical embedding to solely revise the target (*i.e.*, forget and forbidden) speaker's identity.

Experimental results demonstrate that **TruS** effectively suppresses forget speakers' identities without retraining, achieving comparable unlearning performance to existing methods. Crucially, **TruS** is the first to generalize to unseen speakers, extending unlearning beyond the training set to block the generation of voices that mimic individuals. Our contributions are summarized:

- We propose **TruS**, the first training-free unlearning method for zero-shot TTS that constrains models from generating both seen and unseen speaker identities.
- We design a dynamic steering mechanism to selectively control identity-related activations with ID-prototype.
- **TruS** demonstrates comparable performance to tremendously trained baselines, and it further supports opt-out and sequential unlearning requests in a scalable manner.

[*]Equal contribution. [†]Corresponding author.

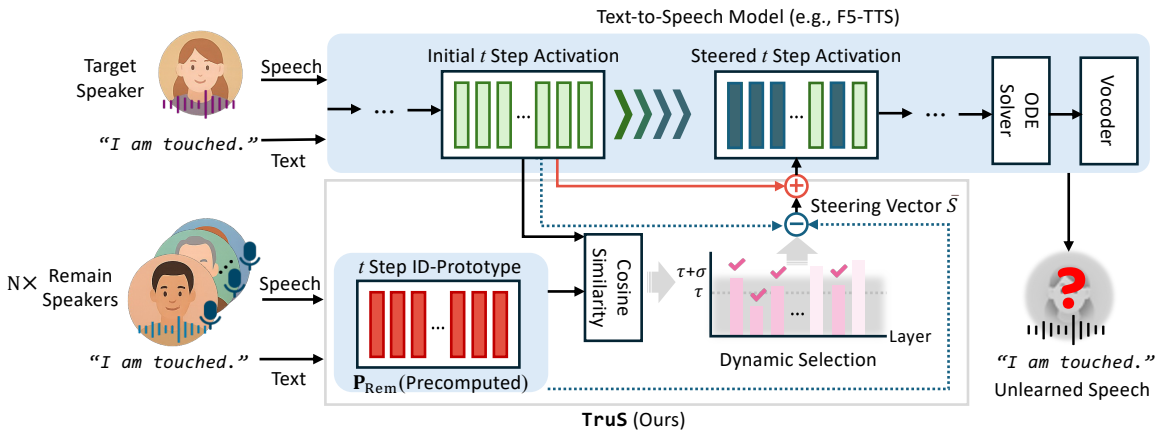


Fig. 2: The overall framework of **TruS**, working with TTS models at inference time. Feature activations at layers and generation steps are optionally steered based on the dynamically selective threshold. With only a single utterance example of a target who wants to be forgotten, our method controls to suppress the identity-related activations without additional training.

2. METHOD

2.1. Motivation and problem formulation

Unlearning in TTS [15] aims to remove the undemanded speaker data in the training set with an efficient training methodology (*i.e.*, without full-tuning). However, as the generalized capability of modern TTS models is increased, the models impose a well-formed speaker embedding space. Consequently, we speculate that removing specific speaker data during training does not guarantee the complete elimination of information. This limitation highlights the necessity of manipulating the model's internal representations to effectively suppress targeted speaker identities. Furthermore, if the request for a new speaker is forgotten, existing works require additional training. Instead, motivated by prior works [18, 19] in natural language processing, we consider that steering activations in TTS models could be an effective and efficient solution to control the speaker identity.

Our new system, **TruS**, enables opt-out unlearning by maintaining a query pool of speaker embeddings for individuals who request their voice not to be synthesized, defined as forget and forbidden sets for training and test data, respectively. At inference, the input speaker embedding from the utterance prompt is compared against the query pool. If a match is found, our lightweight steering module controls hidden representations away from the target identity.

Our **TruS** is built upon the recent fabulous TTS models, such as F5-TTS [5], VoiceBox [3]. To illustrate our method concretely, we adopt F5-TTS as a baseline due to its strong performance and recent adoption as a representative flow-matching TTS model. In the following, we describe our approach on top of this baseline for clarity, though the method is generally applicable to intermediate blocks of other TTS architectures as well. We note that our method is model-agnostic, easily wrapping around existing TTS pipelines.

2.2. Identity-specific steering vector

To suppress the identity-related feature of the target (*i.e.*, forget or forbidden) speaker in the generated speech, **TruS** steer hidden activations with dynamic selection of salient layers. Existing works [20, 21, 19] for steering in LLMs often require a lot of hyperparameters to control the tradeoff between the fidelity to the prompt and the generation quality. Our **TruS** overcomes this challenge through dynamic steering with only one-shot reference example.

Specifically, we prebuild a prototypical identity vector, shortly *ID-prototype*, at each DiT [22] block layer with N remain speakers' utterances. In particular, the outputs of FFN in DiT blocks contain strong timbre and identity signals after non-linear channel mixing [23]. This design allows us to capture identity-specific differences without degrading intelligibility or prosody. The extracted hidden activations over the remain speakers are averaged to build an ID-prototype, which is our base centroid point.

During inference of TTS models, our **TruS** simultaneously computes steering vectors and controls the hidden activations to remove the target speaker's identity. Given a forget utterance U_{For}, each steering vector is independently defined by activations extracted from all L transformer layers: $X_{\text{For}}^{(\ell)} = \text{FFN}^{(\ell)}(U_{\text{For}})$, where $\ell \in \{1, \ldots, L\}$. The ID-prototype is obtained with activations:

$$\mathbf{P}_{\text{Rem}}^{(\ell)} = \sum\nolimits_{N} X_{\text{Rem}}^{(\ell)}/N. \tag{1}$$

Our identity-specific steering vector $S^{(l)}$ for each target speaker is defined as the difference of those activations at every block:

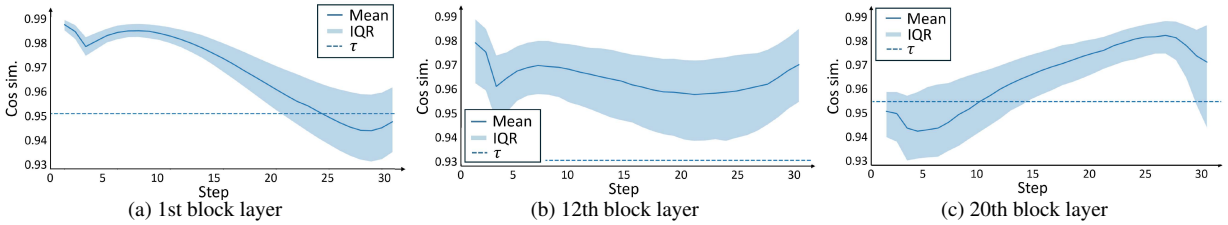
$$S^{(l)} = X_{\text{For}}^{(l)} - \mathbf{P}_{\text{Rem}}^{(\ell)} \tag{2}$$

The ℓ_2-normalized steering vectors, $\bar{S}^{(\ell)}$ represents the identity-related direction of the target speaker within the latent space. They form the basis for suppressing identity-related representation while preserving linguistic content and paralinguistic attributes.

2.3. Dynamic selection of steering layers

Even though the identity-specific cues are encoded to build the basis of activation steering, we argue that *'not all layers contribute equally to maintain speaker identity'*. As illustrated in Fig. 3, cosine similarities between $\mathbf{P}_{\text{Rem}}^{(\ell)}$ and $X_{\text{For}}^{(\ell)}$ are dynamically varying at each generation step. With such a key finding, layers that exhibit lower cosine similarity are considered potential intervention points, since they diverge more strongly between the target speaker and ID-prototype. Compared to our base hook point, FFN, we also empirically found that alternative hook points, such as attention outputs or residual streams, often interfere more strongly with linguistic alignment or prosodic stability.

To dynamically automate layer selection, we propose a dynamic ID threshold τ grounded in global layer statistics. The step-wise similarity $c^{(l,t)}$ is given by the cosine similarity of prototype $\mathbf{P}_{\text{Rem}}^{(\ell)}$



(a) 1st block layer (b) 12th block layer (c) 20th block layer

Fig. 3: Examples of step-wise cosine similarities between hidden activations of target speaker and ID-prototype, at 1st, 12th, 20th layers. Similarity increases in later steps at later layer, and decreases in earlier ones, indicating the need for dynamic layer- and step-specific steering.

Methods	Training hours	WER-R↓	SIM-R↑	WER-F↓	SIM-F↓	spk-ZRF-R	spk-ZRF-F↑
F5-TTS [5]	-	1.95	0.678	3.36	0.657	0.908	0.925
Finetuning	52	2.07	0.654	3.13	0.656	0.911	0.924
SGU [15]	145	2.12	0.290	3.70	0.106	0.935	0.959
TGU [15]	474	2.67	0.532	3.96	0.648	0.914	0.925
TruS	0	1.95†	0.678†	3.69	0.479	0.908†	0.928

Table 1: Quantitative results on LibriSpeech (-R) and the forget test set (-F) on Emilia. Since our method solely works for the forget set, † scores follow the original model. Training hours are reported based on two A6000 GPUs.

and activation $X_{\text{For}}^{(\ell)}$. To determine the dynamic threshold, the global mean and variance across all L layers and T steps are computed:

$$\mu = \frac{1}{L}\sum_{t=1}^{L} c^{(l,t)}, \quad \sigma^2 = \frac{1}{L}\sum_{l=1}^{L}(c^{(l,t)} - \mu)^2 \tag{3}$$

The threshold $\tau = \mu + k\sigma$ is defined with k that controls the tolerance band, and layers are selected based on their average similarity.

In addition, at the step level, we perform a finer filtering within each selected layer. Let $\theta^{(\ell)}$ denote the mean similarity across all steps in layer ℓ. Only steps t, satisfying $\cos\left(X_{\text{For}}^{(\ell,t)}, \mathbf{P}_{\text{Rem}}^{(\ell,t)}\right) < \theta^{(\ell)}$, are retained as admissible intervention points. This two-stage criterion, consisting of a global thresholding across layers followed by local step-wise filtering, results in each selected layer being associated with a subset of admissible steps. In this way, the method adaptively identifies the specific time–layer regions where the identity gap is most pronounced. By dynamically tailoring both layer and step selection per sample, steering avoids excessive or misplaced interventions while maintaining generation ability.

2.4. Unlearning via inference-time steering

After the steering layers are automatically selected, the activations are modified on-the-fly during the denoising process. For each chosen layer l' and flow step k, the pre-computed steering vector $\bar{S}_{l,k}$ is applied to suppress identity-specific information. In order to remove only the component aligned with the identity-related direction while preserving other linguistic and prosodic content, the projection of the activation onto the steering vector is subtracted:

$$\bar{X}_{\text{For}}^{(\ell',t)} = X_{\text{For}}^{(\ell',t)} - \alpha^{(t)} < X_{\text{For}}^{(\ell',t)}, \bar{S}^{(\ell',t)} > \bar{S}^{(\ell',t)}, \tag{4}$$

where the steering strength $\alpha^{(k)}$ serves as a scaling factor to control the intensity of the intervention over flow steps. This design follows empirical observations in Fig. 3 that early layers diverge more strongly during later steps, whereas late layers diverge more during initial steps. Without a burden of training, **TruS** therefore suppresses the contribution of the target speaker to intermediate representations at inference time with the additional benefit of successfully preserving the naturalness of speech.

3. EXPERIMENTS

3.1. Experimental settings

Datasets. Our baseline is F5-TTS [5] pretrained on Emilia [24], a large-scale multi-lingual corpus. We note that only the English subset is considered in this study. For the standard unlearning evaluation protocol, 10 training-set speakers are designated as the forget set, with 300 seconds held out for testing, while the remaining Emilia speakers form the training set. LibriSpeech test-clean corpus [25] is employed to evaluate both the preservation of generation performance for remain speakers and the generalized unlearning capability. The forbidden set consists of 10 unseen sexually-balanced speakers (∼300 seconds each), and the other speakers belong to remain set. The audio of the remaining speakers was used to evaluate generalization performance. This protocol enables systematic evaluation of whether the method suppresses forbidden speakers while avoiding unintended degradation for unseen voices. Finally, emotional fidelity is assessed on CREMA-D [26], where 10 speakers are selected as a forget set, and 30 utterances per speaker are generated.

Implementation details. Our method entirely works at inference time, operating on a pretrained open-source TTS model, F5-TTS [5] without any additional finetuning. We also reimplement baseline based on authors' released code, including SGU [15] and TGU [15], on F5-TTS. In practice, we empirically set the steering strength to $\alpha = 1.5$ for Emilia and $\alpha = 1.2$ for unseen speakers from LibriSpeech to balance forgetting effectiveness and generation quality.

Metrics. Following prior work [15], we assess the performance with SIM-Spk [3], WER, spk-ZRF [15], and SIM-Emo. SIM-Spk measures the embedding similarity between generated and reference speech, extracted using a pretrained speaker verification model [27]. While WER quantifies linguistic fidelity by comparing the transcription of generated speech with the reference text, Spk-ZRF [15] measures the degree of randomness in the generated voices for forget speakers. Finally, SIM-Emo employs emotion2vec [28] to measure the similarities between emotion embeddings from the generated and reference speech.

Methods	Forget	WER-UF ↓	SIM-UF ↑ ↓	spk-ZRF-UF
F5-TTS [5]	✗	2.03	0.668	0.906
Finetuning	✗	2.18	0.646	0.912
SGU [15]	✗	2.05	0.370	0.930
TGU [15]	✗	2.68	0.513	0.907
TruS	✓	3.31	0.433	0.908

Table 2: Performance on unseen forbidden set (-UF) in LibriSpeech.

Methods	Forget	SIM-Spk↑ ↓	SIM-Emo↑
Unconditional F5-TTS	✗	0.048	0.733
Finetuning	✗	0.230	0.730
TruS	✓	0.123	0.736

Table 3: Performance comparison on CREMA-D. We report cosine similarities based on speaker embedding (SIM-Spk) and emotional embedding (SIM-Emo) with emotion2vec [28].

3.2. Results

Forget speakers. Our evaluation focuses on two aspects: (i) the degree to which the target speaker identity is suppressed in the forget set, quantified by reductions in SIM-F, and (ii) the preservation of linguistic content, reflected by controlled changes in WER-F. Table 1 reports the results on the forget set. Optimization-based unlearning methods (*i.e.*, SGU and TGU) [15] require verification of both remain and forget performance to ensure that retraining does not compromise the generative quality. In contrast, **TruS** does not involve training, but instead applies steering during synthesis only for forget or forbidden speakers. Therefore, generations for remain speakers are identical to those produced by the pretrained checkpoint. The results show that **TruS** achieves a substantial reduction in SIM-F, confirming the effective suppression of the target speaker's identity. At the same time, WER-F score on **TruS** outperforms SGU and TGU, which cost tremendous training, *e.g.*, 145 and 474 hours for 2 A6000 GPUs, respectively. These results indicate that linguistic accuracy is preserved even as speaker similarity is reduced. This asymmetric outcome, strong suppression of identity with minimal impact on content fidelity, provides direct evidence that **TruS** appropriately adjusts hidden representations related to identity rather than introducing indiscriminate perturbations. Meanwhile, SGU shows the highest spk-ZRF-F among the comparison models, while SIM-F drops significantly. However, SIM-R also declines substantially, indicating that it fails to selectively forget only the influence of the forget targets. Moreover, even for the forgotten, SGU shows the highest WER-F among the comparison models, showing the weakness to preserve the generalization performance during unlearning optimization. TGU does not guarantee superb performance nor that its SIM-F will not decrease sufficiently. The decline in SIM-R, together with the increase in WER-R and WER-F, shows that TGU, like SGU, also suffers from degraded generalization performance.

Generalization to unseen forbidden speakers. **TruS** is evaluated on unseen speakers from LibriSpeech [25], which serve as a proxy for the forbidden set, in order to examine whether the unlearning effects extend beyond the training identities. Table 2 presents that while WER-UF slightly degrades, but speaker similarities (SIM-UF) are effectively decreased with **TruS**. We note that in contrast to our approach, all other baselines exhibit high SIM-UF for unseen speakers, since they do not perform unlearning for the forbidden set. Especially, both SGU and TGU degrade the speaker fidelity performance than the original model due to the additional retraining. Practically, the results demonstrate that **TruS** effectively blocks the synthesis of voices mimicking restricted individuals, while maintaining robust performance for original methods, demonstrating the validity of training-free unlearning.

τ	SIM-F↓	WER-F↓	spk-ZFR-F↑	SIM-UF↓	WER-UF↓	spk-ZRF-UF↑
$\mu-\sigma$	0.567	3.51	0.930	0.533	2.38	0.909
μ	0.537	3.66	0.930	0.480	3.87	0.907
$\mu+\sigma$	0.477	3.48	0.933	0.436	3.76	0.909
all	0.461	3.53	0.934	0.424	5.08	0.909

Table 4: Performance over different layer selection strategies.

#	SIM-F↓	WER-F↓	spk-ZRF-F↑	SIM-UF↓	WER-UF↓	spk-ZRF-UF↑
N=10	0.526	3.43	0.932	0.487	5.53	0.909
N=30	0.477	3.48	0.933	0.436	3.76	0.909
N=50	0.509	3.60	0.933	0.475	2.71	0.911

Table 5: Performance over different numbers of remain speakers.

Emotion preservation. We measure SIM-Spk and SIM-Emo using the emotional speech dataset [26] to verify whether the emotion attributes are preserved even after applying **TruS** on a zero-shot setting (*i.e.*, all methods not trained on CREMA-D). The results are presented in Table 3. Surprisingly, **TruS** obtains a low SIM-Spk score, which demonstrates that the target speaker has been successfully forgotten. To establish the lower bound of SIM-Emo, we also generate voices using F5-TTS without conditioning on a speech prompt. In contrast, 'Finetuning' technique yields lower SIM-Emo than 'Unconditional F5-TTS', suggesting that finetuning negatively affects the preservation of emotion. Meanwhile, **TruS** achieves higher SIM-Emo than both baselines, indicating that it retains emotion attributes relative to the reference audio even after unlearning. Therefore, **TruS** performs effective unlearning (or prohibition) while maintaining attribute preservation beyond speaker identity.

3.3. Ablation study

Layer filtering criteria. We examine how different layer filtering criteria affect unlearning by steering only layers where their similarity is below '$\mu-\sigma$', 'μ', '$\mu+\sigma$', or all layers. Table 4 shows that the '$\mu+\sigma$' criterion provides the most reliable performance over all metrics, whereas steering all layers achieves slightly lower SIM-F/UF but substantially degrades WER-F/UF. Additionally, more strict thresholds ('μ' or '$\mu-\sigma$') yield insufficient suppression for identity removal. These results indicate that '$\mu+\sigma$' effectively isolates layers that capture identity-specific signals while avoiding disruption of phonetic fidelity.

The pool size of ID-prototype. Table 5 examines the performance varying the retain-speaker pool size N=$\{10, 30, 50\}$ under the '$\mu+\sigma$' criterion. With only 10 speakers, suppression is relatively weak and WER is unstable. At $N = 30$, SIM-F/UF reaches the lowest values with stable WER, offering the most balanced trade-off between suppression and fidelity. Increasing the pool size to 50 further improves text fidelity, but SIM scores are slightly increased. Hence, a pool size of around 30 provides the most appropriate balance between the effectiveness of suppression and generation quality.

4. CONCLUSION

We introduce **TruS**, a training-free framework for opt-out speaker unlearning in zero-shot TTS for the first time. **TruS** compares the hidden activations of the target speaker desired to unlearn against an averaged ID-prototype embedding, then steers the identity basis of activation. We believe this paradigm shift offers a practical foundation for future generative speech systems, establishing a scalable solution to user-driven privacy requests.

5. REFERENCES

[1] F. Lux, J. Koch, and N. T. Vu, "Low-resource multilingual and zero-shot multispeaker tts," 2022, pp. 741–751.

[2] S. Chen, Y. Feng, L. He, T. He, W. He, Y. Hu, B. Lin, Y. Lin, Y. Pan, P. Tan, et al., "Takin: A cohort of superior quality zero-shot speech generation models," *arXiv preprint arXiv:2409.12139*, 2024.

[3] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, et al., "Voicebox: Text-guided multilingual universal speech generation at scale," in *NeurIPS*, 2023.

[4] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan, et al., "E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts," in *SLT Workshop*, 2024.

[5] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," *CoRR*, 2024.

[6] A. Alali and G. Theodorakopoulos, "Partial fake speech attacks in the real world using deepfake audio," *Journal of Cybersecurity and Privacy*, vol. 5, no. 1, pp. 6, 2025.

[7] N. Tomashenko, X. Miao, E. Vincent, and J. Yamagishi, "The first voiceprivacy attacker challenge," in *ICASSP*. IEEE, 2025.

[8] L. Juvela and X. Wang, "Collaborative watermarking for adversarial speech synthesis," in *ICASSP*, 2024.

[9] W. Zong, Y.-W. Chow, W. Susilo, J. Baek, and S. Camtepe, "{AudioMarkNet}: Audio watermarking for deepfake speech detection," 2025, pp. 4663–4682.

[10] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot, "Machine unlearning," in *SP*, 2021.

[11] T. Baumhauer, P. Schöttle, and M. Zeppelzauer, "Machine unlearning: Linear filtration for logit-based classifiers," *Machine Learning*, vol. 111, no. 9, pp. 3203–3226, 2022.

[12] A. Warnecke, L. Pirch, C. Wressnegger, and K. Rieck, "Machine unlearning of features and labels," in *NDSS*. Internet Society, 2023.

[13] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli, "Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher," in *AAAI*, 2023.

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

[15] T. Kim, J. Kim, D. Kim, J. H. Ko, and G.-M. Park, "Do not mimic my voice: Speaker identity unlearning for zero-shot text-to-speech," in *ICML*, 2025.

[16] J. Seo, S.-H. Lee, T.-Y. Lee, S. Moon, and G.-M. Park, "Generative unlearning for any identity," in *CVPR*, 2024.

[17] T. Xie, S. Yang, C. Li, D. Yu, and L. Liu, "Emosteer-tts: Fine-grained and training-free emotion-controllable text-to-speech via activation steering," *arXiv preprint arXiv:2508.03543*, 2025.

[18] N. Rimsky, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner, "Steering llama 2 via contrastive activation addition," in *ACL*, 2024, pp. 15504–15522.

[19] A. M. Turner, L. Thiergart, G. Leech, D. Udell, J. J. Vazquez, U. Mini, and M. MacDiarmid, "Steering language models with activation engineering," *arXiv preprint arXiv:2308.10248*, 2023.

[20] Y. Chen, H. Jhamtani, S. Sharma, C. Fan, and C. Wang, "Steering large language models between code execution and textual reasoning," in *ICLR*, 2025.

[21] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[22] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *ICCV*, 2023.

[23] T.-Q. Lin, H.-C. Cheng, H.-y. Lee, and H. Tang, "Identifying speaker information in feed-forward layers of self-supervised speech transformers," *arXiv preprint arXiv:2506.21712*, 2025.

[24] H. He, Z. Shang, C. Wang, et al., "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," in *SLT*, 2024.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP*, 2015.

[26] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. on Aff. Comp.*, vol. 5, no. 4, pp. 377–390, 2014.

[27] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech*, 2020.

[28] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," in *ACL*, 2024, pp. 15747–15760.