ICML 2025

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

Taesoo Kim* [1 2]

Jinju Kim* [1 3]
(Presenter)

Dong Chan Kim[1]

Jong Hwan Ko† [1]

Gyeong-Moon Park † [4]

* Equal Contribution
† Corresponding Author
[1] Sungkyunkwan University
[2] KT Corporation
[3] Carnegie Mellon University
[4] Korea University

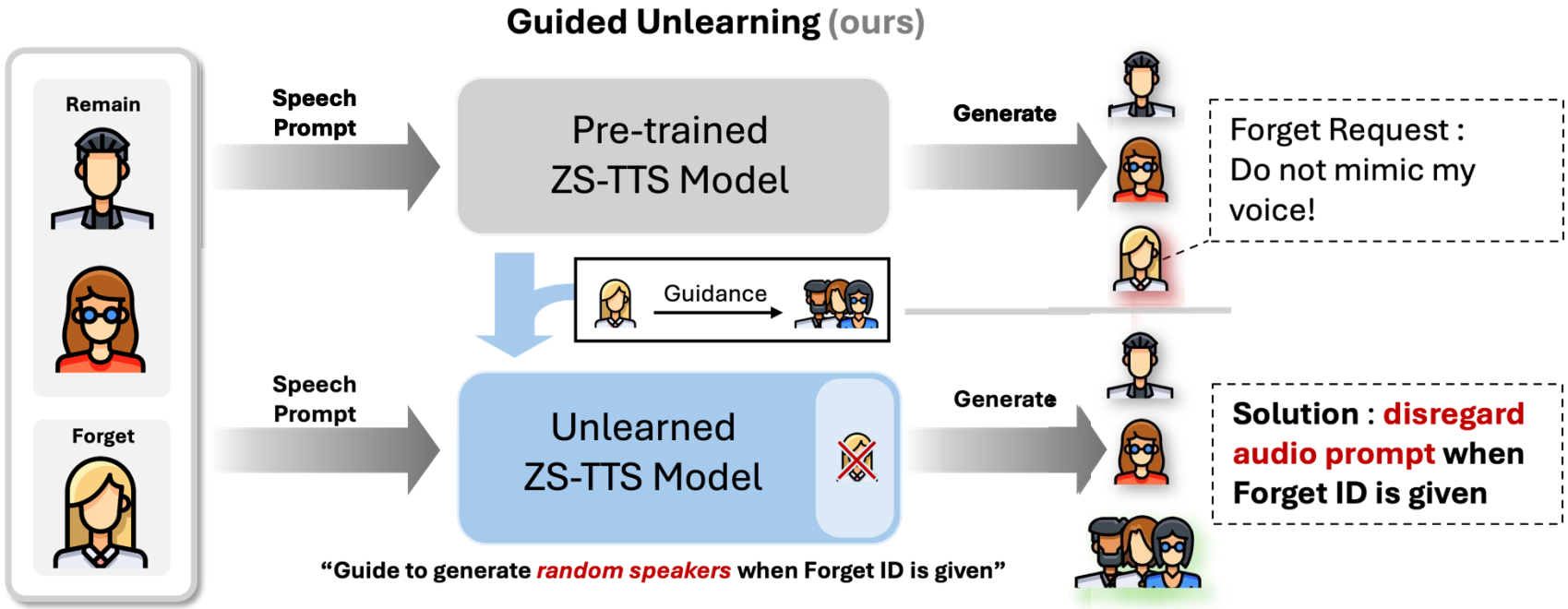
Motivation

- **Privacy Hazard of Zero-Shot Text-to-Speech (ZS-TTS)**:
  - ZS-TTS models can replicate a person’s voice using only a few seconds of audio, making it easy to impersonate someone without consent.
  - A person's voice is a **biometric identifier**, voice cloning poses severe privacy risks.

- **Zero-Shot Generalization Poses Challenges in Unlearning**:
  - ZS-TTS models can generalize to new voices not seen in training, making **traditional unlearning insufficient**.
  - Effective unlearning requires actively neutralizing speaker style even in **unseen conditions**.

Jinju Kim (perla0328@g.skku.edu)

Overview



Jinju Kim (perla0328@g.skku.edu)

Demo

Forget Speaker
Ground Truth

Pre-trained
ZS-TTS Model

Cloned with ZS-TTS

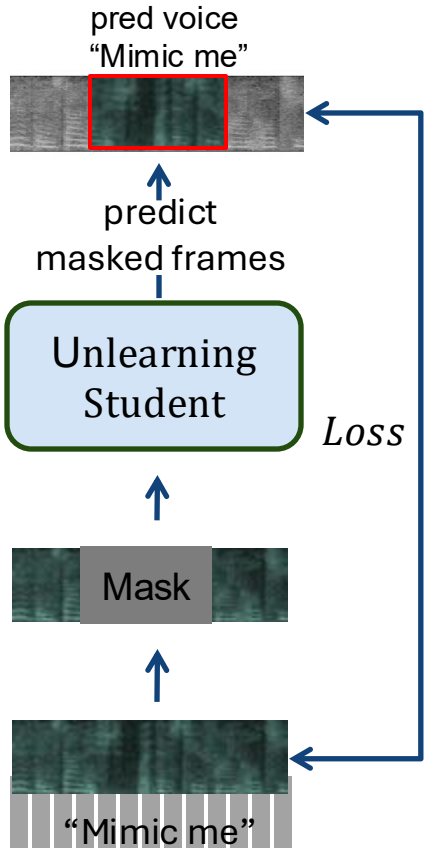
Guided Unlearning (ours)

Unlearned
ZS-TTS Model

TGU unlearned (OURS)

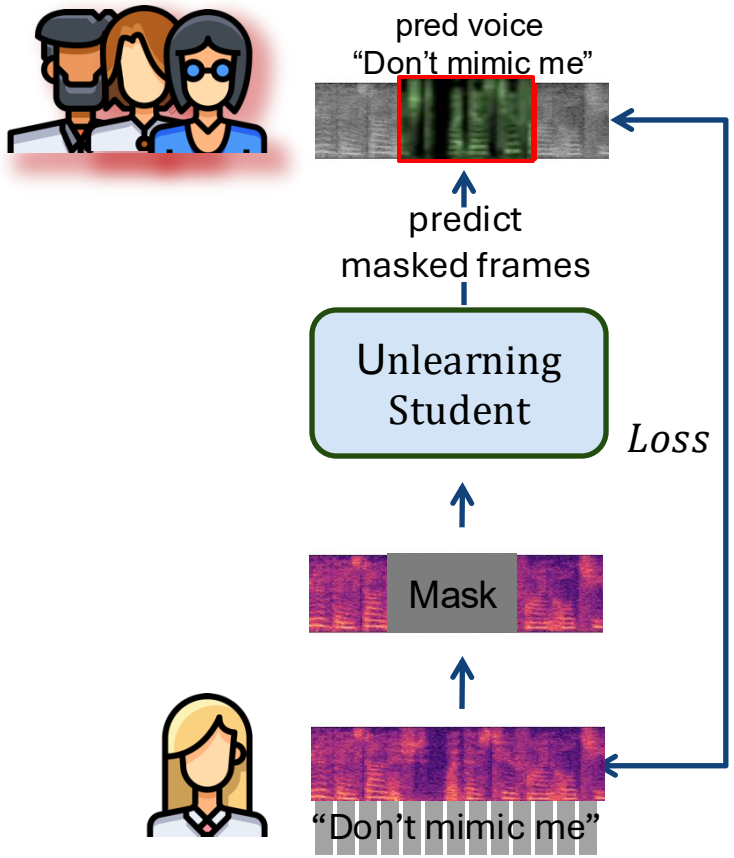
Jinju Kim (perla0328@g.skku.edu)

Challenge



Jinju Kim (perla0328@g.skku.edu)

Challenge

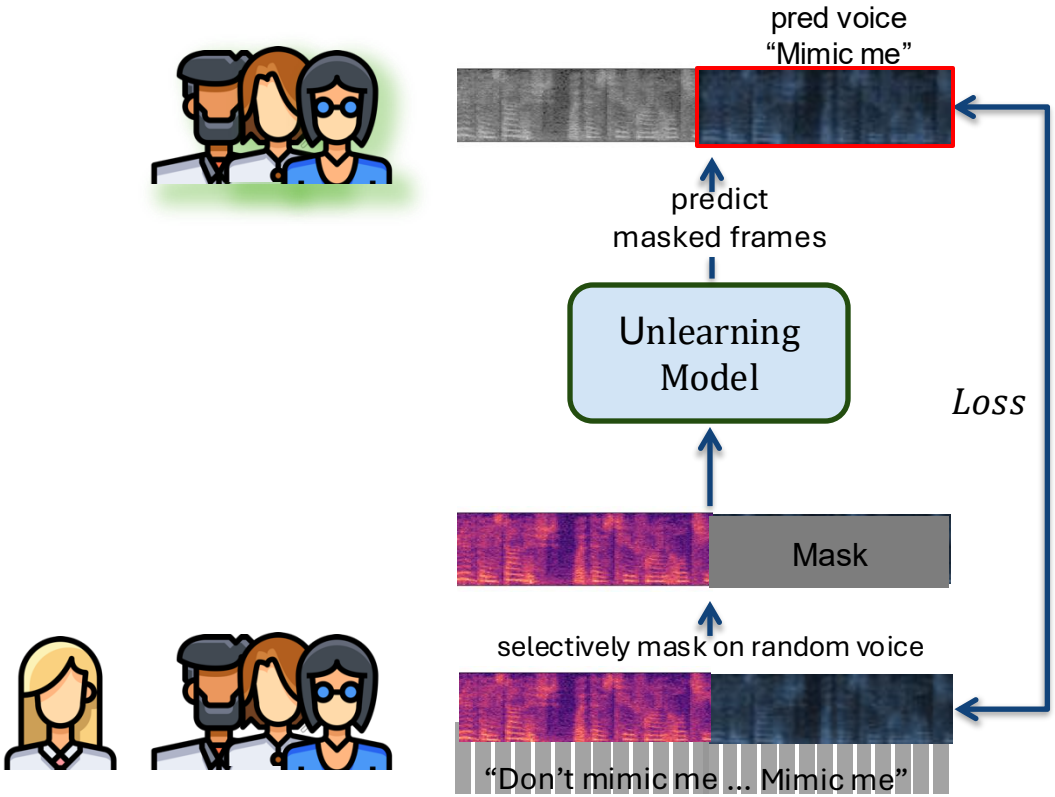


Aligned voice pairs do not naturally exist !

Jinju Kim (perla0328@g.skku.edu)

Methods

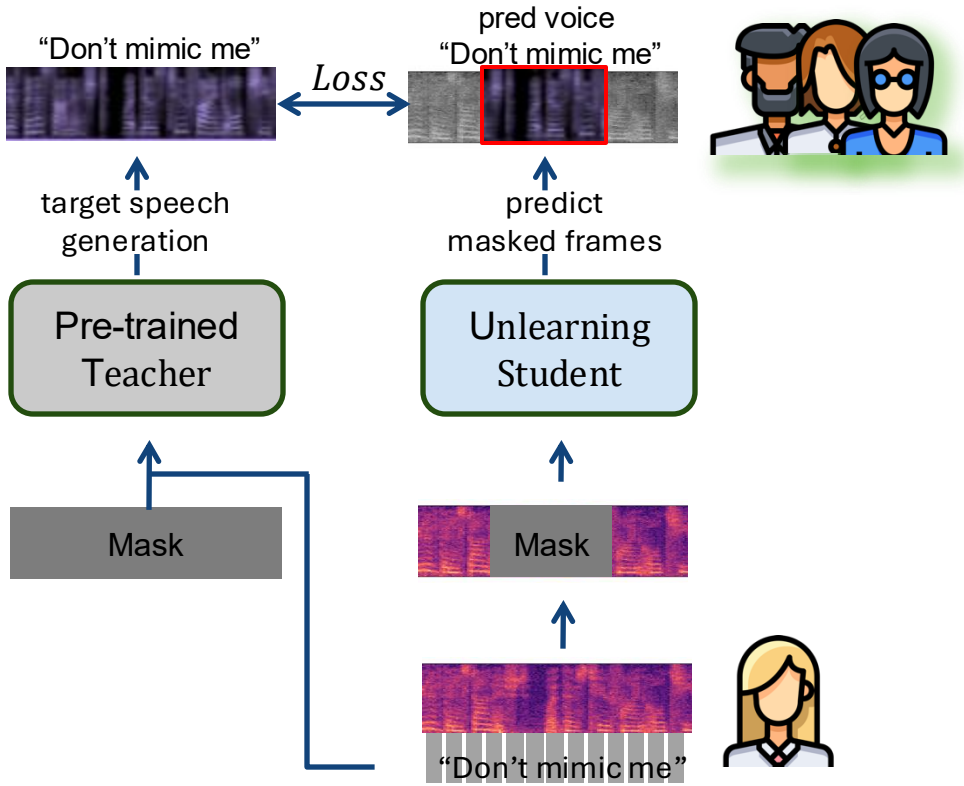
Sample-Guided Unlearning (SGU)



Jinju Kim (perla0328@g.skku.edu)

Methods

Teacher-Guided Unlearning (TGU)



Jinju Kim (perla0328@g.skku.edu)

Methods

speaker-Zero Retrain Forgetting (spk-ZRF)

- **Limitations of Conventional Evaluation Methods**
  - focus on measuring performance difference between forget and remain sets
  - does not indicate whether the model has **truly forgotten**

"Don't mimic me"

Unlearned
ZS-TTS Model

"Don't mimic me"

↕

purely randomly generated samples

Jinju Kim (perla0328@g.skku.edu)

Experiment

Forgetting In-Domain Speakers

Methods	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↓	spk-ZRF-R	spk-ZRF-F ↑
Original◇	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	-	-	-

Table 1. Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set (-F).

Jinju Kim (perla0328@g.skku.edu)

Experiment

Forgetting In-Domain Speakers

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	-	-	-

Table 1. Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set (-F).

Jinju Kim (perla0328@g.skku.edu)

Experiment

Forgetting In-Domain Speakers

Methods	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↓	spk-ZRF-R	spk-ZRF-F ↑
Original◇	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	-	-	-

Table 1. Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set (-F).

Jinju Kim (perla0328@g.skku.edu)

Experiment

Forgetting In-Domain Speakers

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	-	-	-

Table 1. Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set (-F).

Jinju Kim (perla0328@g.skku.edu)

Experiment

Forgetting In-Domain Speakers

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	-	-	-

Table 1. Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set (-F).

Jinju Kim (perla0328@g.skku.edu)

Experiment

Forgetting In-Domain Speakers

Methods	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↓	spk-ZRF-R	spk-ZRF-F ↑
Original◇	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	-	-	-

Table 1. Quantitative results on LibriSpeech test-clean evaluation set (-R) and the forget evaluation set (-F).

Jinju Kim (perla0328@g.skku.edu)

Experiment

Scalability

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 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.

Jinju Kim (perla0328@g.skku.edu)

Experiment

Unlearning Out-of-Domain Speakers

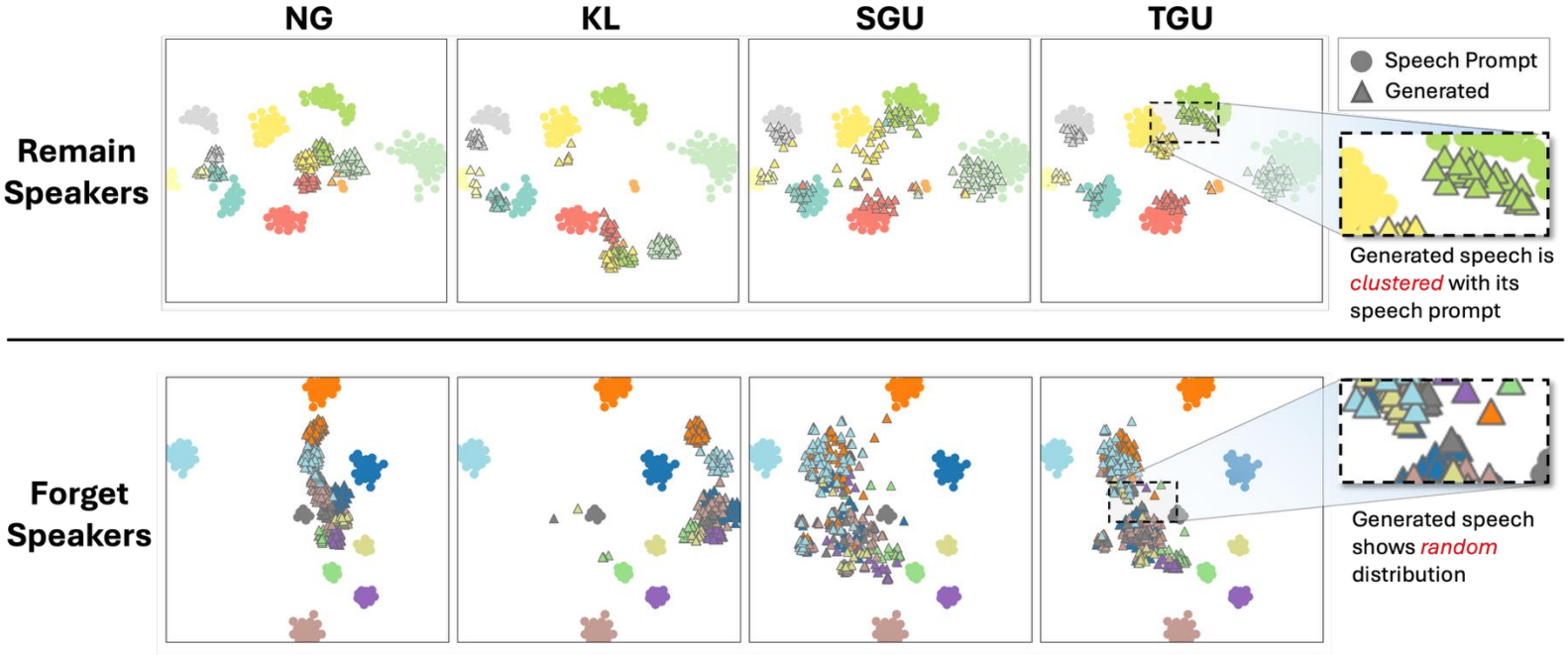
Methods	WER-R ↓	SIM-R ↑	WER-F ↓	SIM-F ↓
Original	2.7	0.649	5.1	0.678
SGU	2.9	0.602	5.5	**0.157**
TGU	**2.5**	**0.630**	**5.3**	0.186
Ground Truth	2.2	-	5.9	-

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

Jinju Kim (perla0328@g.skku.edu)

Experiment

Speaker Analysis



Jinju Kim (perla0328@g.skku.edu)

Conclusion

- **First application** of machine unlearning for Zero-Shot TTS

- **Teacher-Guided Unlearning (TGU)**
  - Injects randomness to erase model's ability to process *forget speaker* audio prompts
  - Prevents unwanted voice replication while preserving original ZS-TTS performance
  - Only **2.6% drop** in speaker similarity (SIM) for *remain speakers*
  - Maintains word error rate (WER) compared to original model

- **speaker-Zero Retrain Forgetting (spk-ZRF)**
  - Measures **randomness** in generated speech
  - Evaluates resistance to **reverse engineering attacks** that could reveal speaker identity

Jinju Kim (perla0328@g.skku.edu)

- END -

Taesoo Kim* [1 2], **Jinju Kim*** [1 3], Dong Chan Kim[1], Jong Hwan Ko† [1], Gyeong-Moon Park† [4]

* Equal Contribution
† Corresponding Author
[1] Sungkyunkwan University
[2] KT Corporation
[3] Carnegie Mellon University
[4] Korea University

presenter contacts
email | perla0328@g.skku.edu
phone | +82) 10-9430-6636