ICML | 2022

Guided-TTS: A Diffusion Model for Text-to-Speech via Classifier Guidance

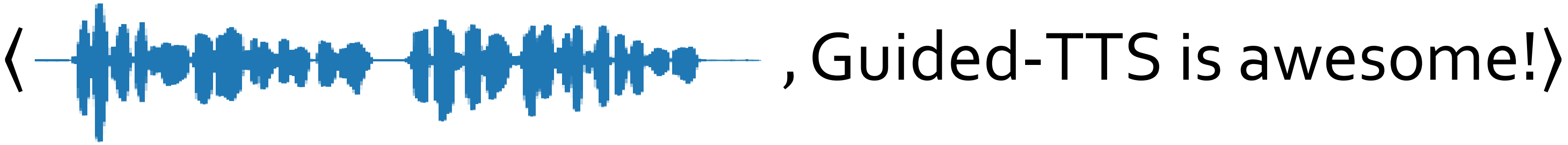
Heeseung Kim*, Sungwon Kim*, Sungroh Yoon

*Equal Contribution

Data Science & Artificial Intelligence Laboratory
Electrical and Computer Engineering
Seoul National University

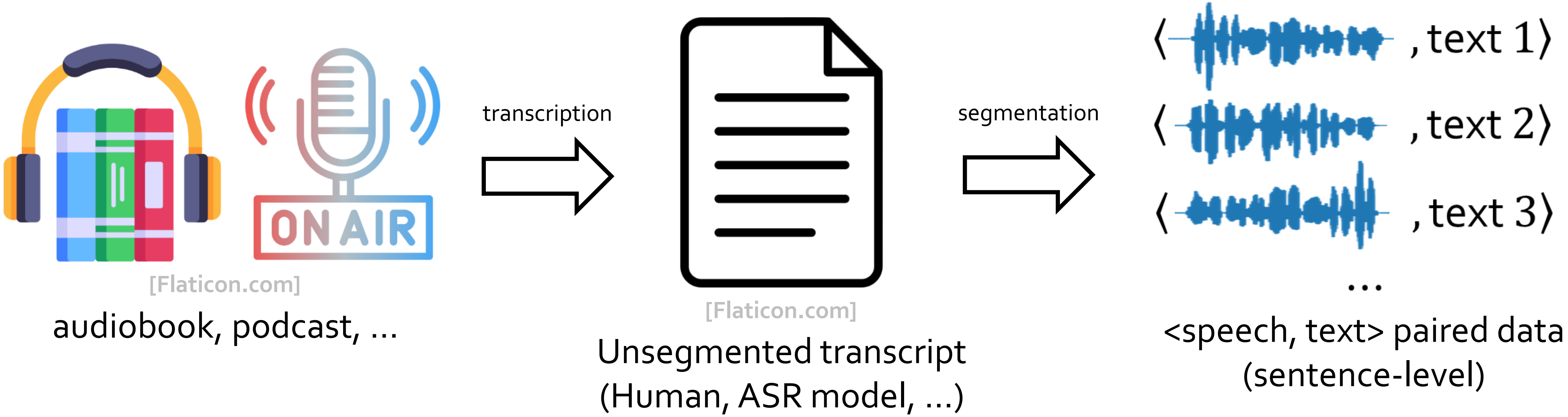
Motivation

- Training data for existing TTS models



⟨ , Guided-TTS is awesome!⟩

- How to train a TTS model with **long-form untranscribed data**?



audiobook, podcast, …

Unsegmented transcript
(Human, ASR model, …)

<speech, text> paired data
(sentence-level)

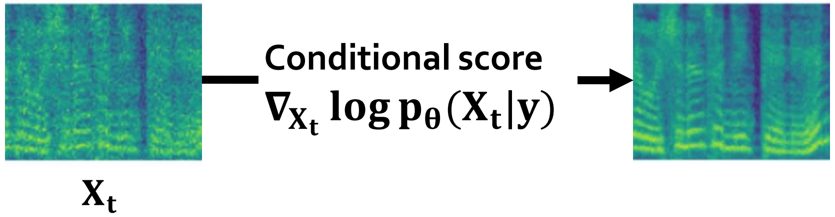
- **Guided-TTS directly uses untranscribed data of the target speaker for training**

Overview of Guided-TTS

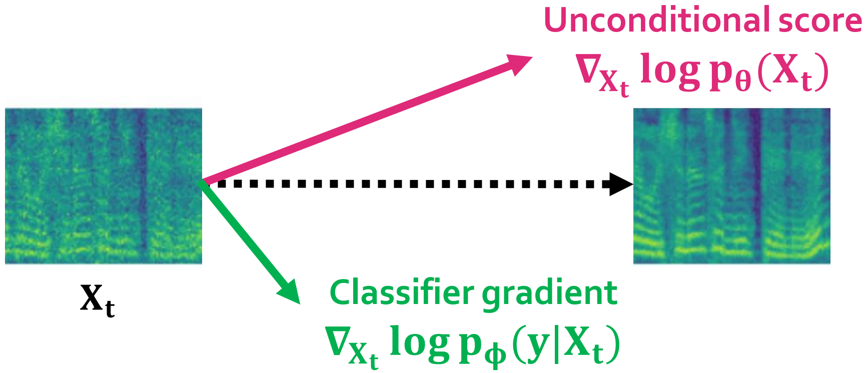
- Guided-TTS = **unconditional DDPM** + **phoneme classifier**

Classifier-guidance $\nabla_{X_t} \log p(X_t|y) = \nabla_{X_t} \log p(X_t) + \nabla_{X_t} \log p(y|X_t)$

X_t : noisy mel-spectrogram
y : phoneme sequence



Existing Diffusion Model for Text-to-Speech

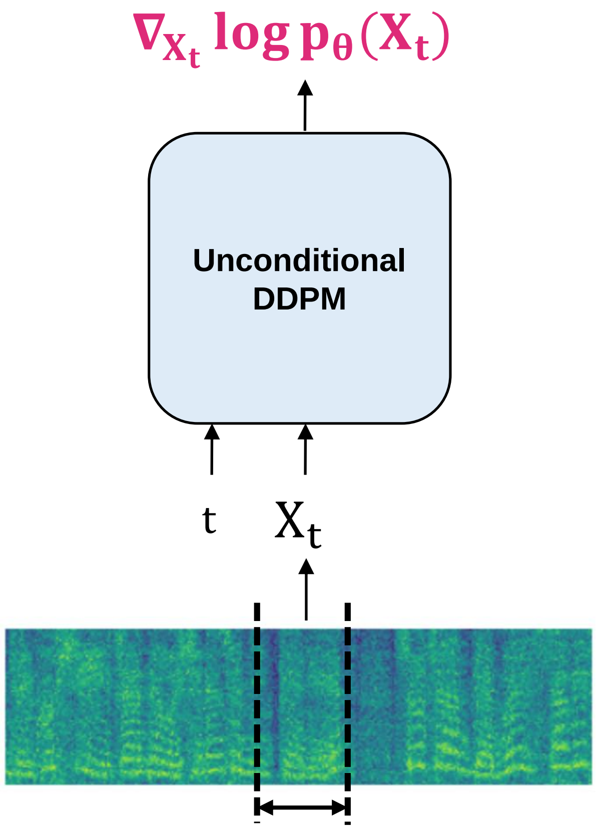

Unconditional score
$\nabla_{X_t} \log p_\theta(X_t)$
X_t
Classifier gradient
$\nabla_{X_t} \log p_\phi(y|X_t)$

Guided-TTS: Text-to-Speech via Classifier Guidance

- Guided-TTS generalizes well to diverse untranscribed datasets with the **single phoneme classifier trained on a large-scale multi-speaker ASR dataset**

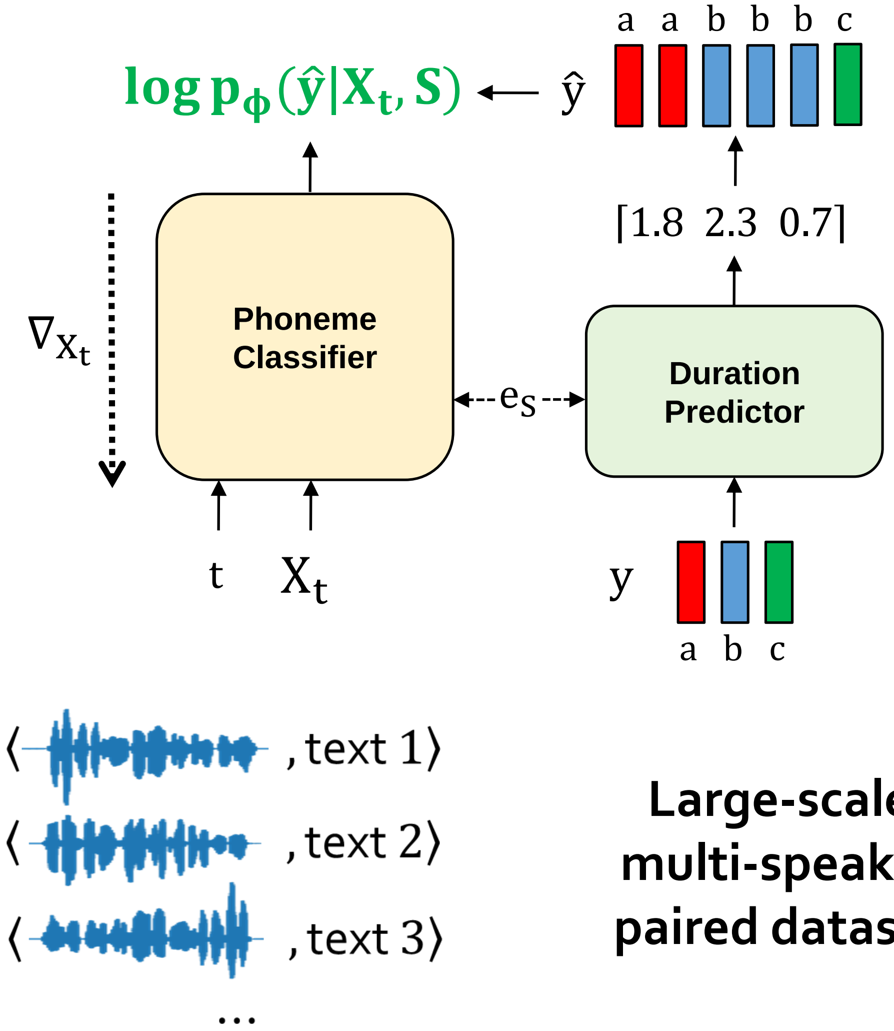
Overview of Guided-TTS

Modeling unconditional score $\nabla_{X_t} \log p_\theta(X_t)$



5-second-long random chunks of untranscribed speech

Modeling classifier gradient $\nabla_{X_t} \log p_\phi(\hat{y}|X_t, S)$

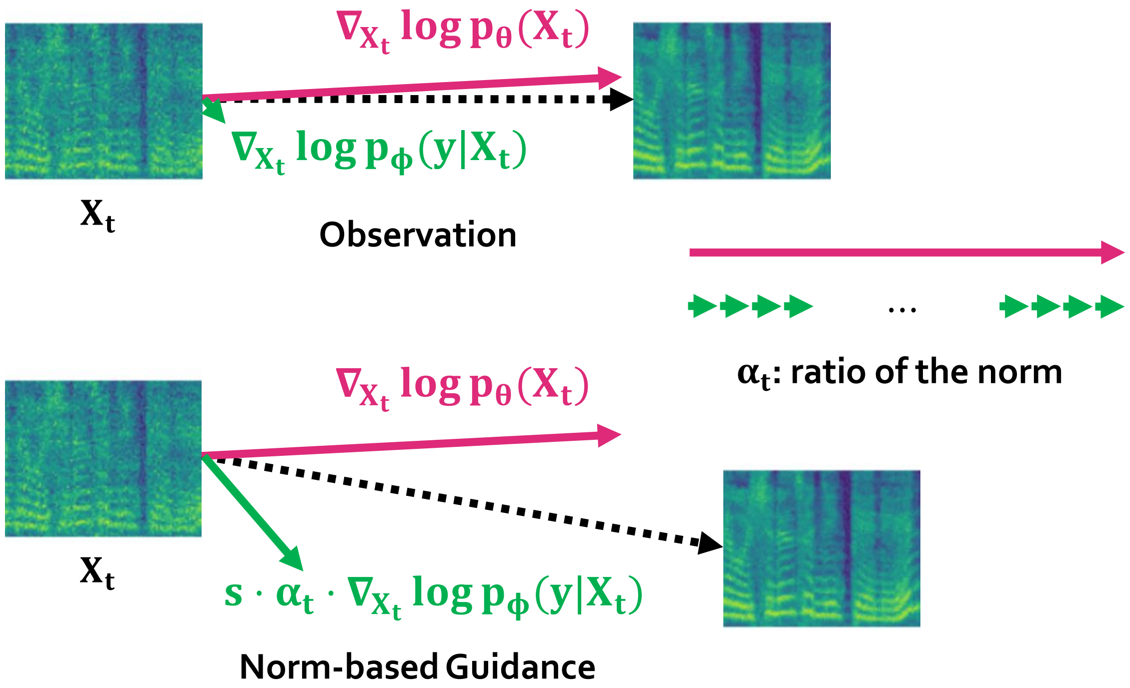


Large-scale multi-speaker paired dataset

Norm-based Classifier Guidance

- Observation
  - As t → 0, ||classifier gradient|| << ||unconditional score|| → **Pronunciation errors**

- Norm-based Guidance: classifier gradient *= Norm-ratio $\left(= \dfrac{\|\nabla_{X_t} \log p_\theta(X_t)\|}{\|\nabla_{X_t} \log p_\phi(y|X_t)\|}\right)$



$\nabla_{X_t} \log p_\theta(X_t)$

$\nabla_{X_t} \log p_\phi(y|X_t)$

X_t

Observation

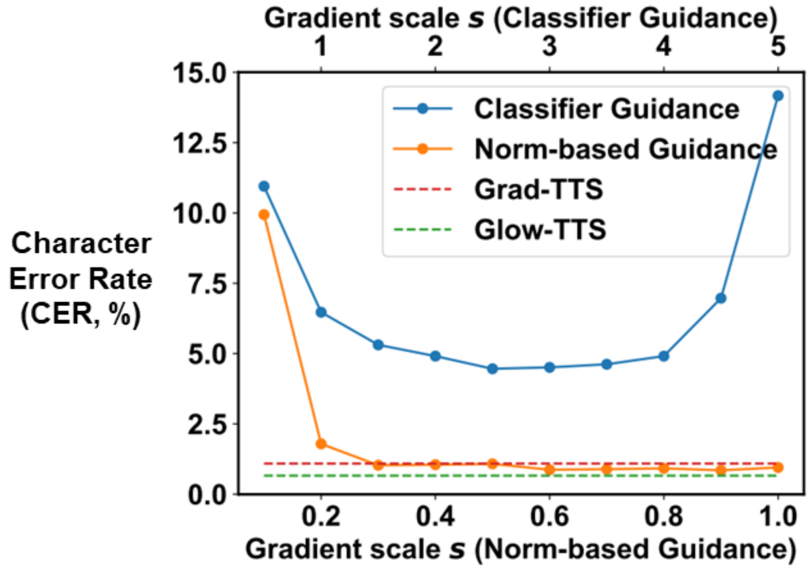
α_t: ratio of the norm

$\nabla_{X_t} \log p_\theta(X_t)$

X_t

$s \cdot \alpha_t \cdot \nabla_{X_t} \log p_\phi(y|X_t)$

Norm-based Guidance

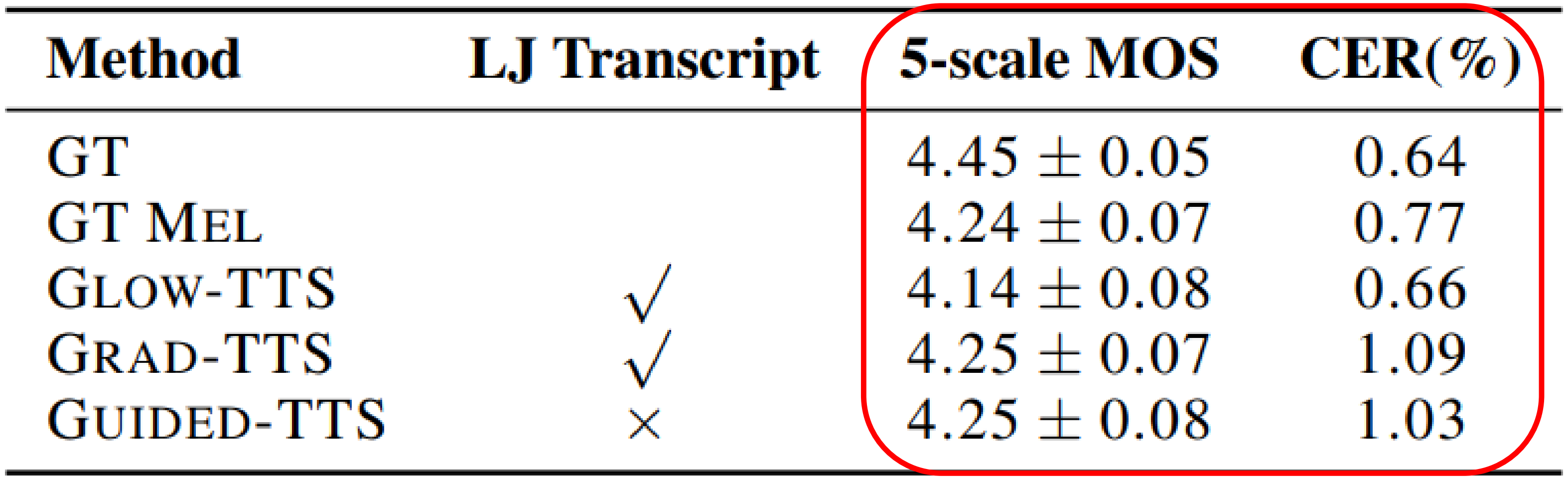


Norm-based Guidance > Classifier Guidance

Song et al., 2021,
Dhariwal et al., 2021

Results (1)

- Comparison with **high-quality TTS models that require target speaker's transcript**



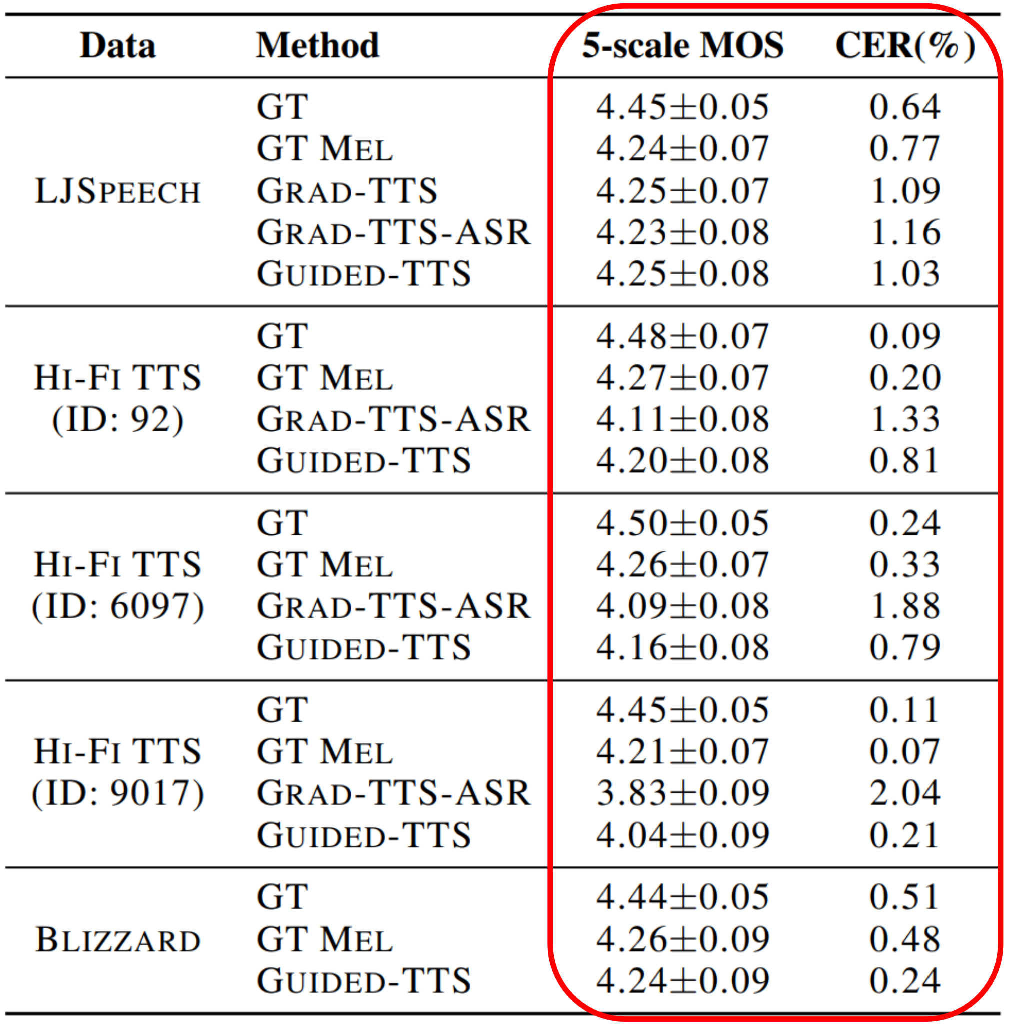
Method	LJ Transcript	5-scale MOS	CER(%)
GT		4.45 ± 0.05	0.64
GT MEL		4.24 ± 0.07	0.77
GLOW-TTS	√	4.14 ± 0.08	0.66
GRAD-TTS	√	4.25 ± 0.07	1.09
GUIDED-TTS	×	4.25 ± 0.08	1.03

Guided-TTS ≈ Grad-TTS > Glow-TTS

Popov et al., 2021 Kim et al., 2020

Results (2)

- Comparison with **Grad-TTS-ASR (construct paired data using pre-trained ASR model)**



Data	Method	5-scale MOS	CER(%)
LJSPEECH	GT	4.45±0.05	0.64
	GT MEL	4.24±0.07	0.77
	GRAD-TTS	4.25±0.07	1.09
	GRAD-TTS-ASR	4.23±0.08	1.16
	GUIDED-TTS	4.25±0.08	1.03
HI-FI TTS (ID: 92)	GT	4.48±0.07	0.09
	GT MEL	4.27±0.07	0.20
	GRAD-TTS-ASR	4.11±0.08	1.33
	GUIDED-TTS	4.20±0.08	0.81
HI-FI TTS (ID: 6097)	GT	4.50±0.05	0.24
	GT MEL	4.26±0.07	0.33
	GRAD-TTS-ASR	4.09±0.08	1.88
	GUIDED-TTS	4.16±0.08	0.79
HI-FI TTS (ID: 9017)	GT	4.45±0.05	0.11
	GT MEL	4.21±0.07	0.07
	GRAD-TTS-ASR	3.83±0.09	2.04
	GUIDED-TTS	4.04±0.09	0.21
BLIZZARD	GT	4.44±0.05	0.51
	GT MEL	4.26±0.09	0.48
	GUIDED-TTS	4.24±0.09	0.24

Guided-TTS > Grad-TTS-ASR

Generalize well to diverse datasets

Conclusion

- Guided-TTS is a new type of TTS model that generates speech given transcript by guiding the unconditional diffusion-based model for speech.

- To the best of our knowledge, Guided-TTS is the first TTS model to leverage the unconditional generative model for speech.

Poster Session 2
Wed 20 Jul 18:30 – 20:30
Hall E #116

Paper

Demo Page

Thank You!