이화여자대학교

DurMI: Duration loss as a membership signal in TTS Models

Saeyeon Hong 252AIG20

EWHA,
THE FUTURE
WE CREATE

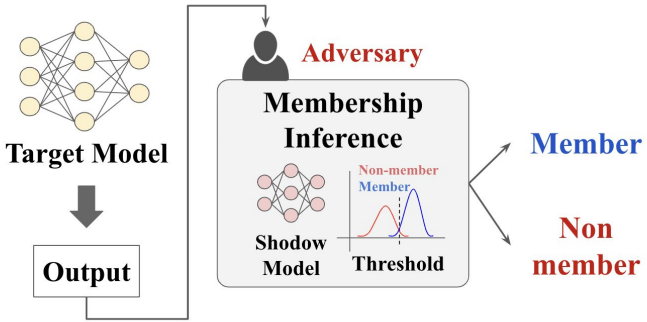
Overview

DurMI: Duration loss as a membership signal in TTS Models

- Membership inference attack in Text-to-Speech (TTS) models.
- Currently under review (ICLR)

Introduction

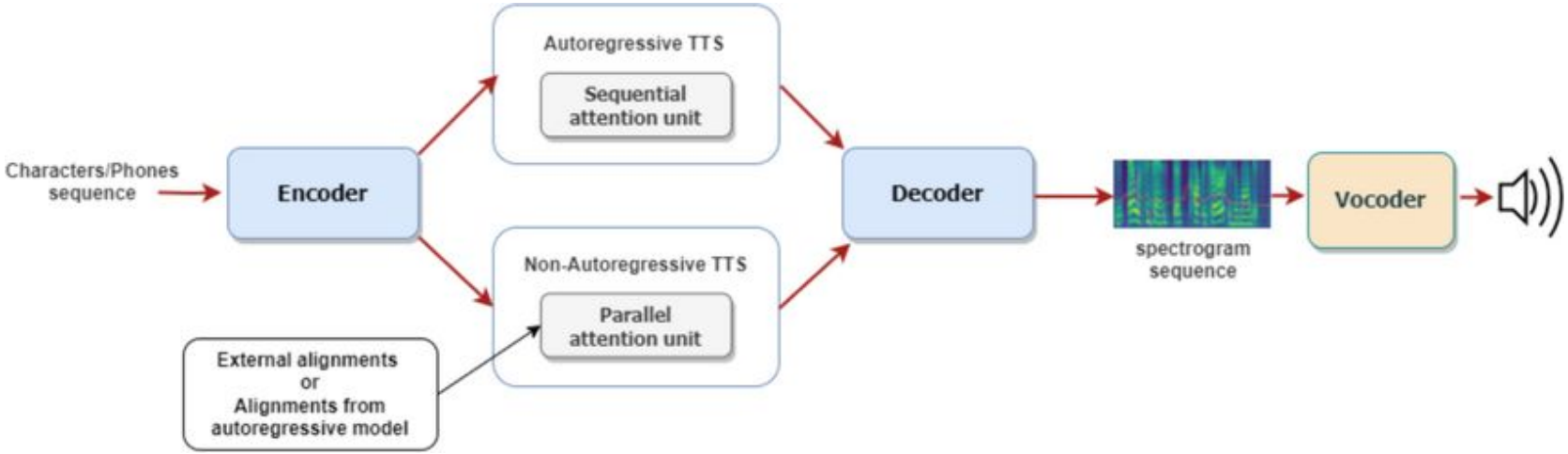
- **Membership Inference Attack (MIA)**: The adversary seeks to determine whether a given input was used during model training.



- While MIA has been extensively explored in computer vision and natural language processing, its application to **Text-to-Speech (TTS) remains underexplored**.
- Comprehensive experiments across two diffusion-based TTS models and three benchmark datasets demonstrate that DurMI significantly outperforms prior diffusion-based MIA methods.

Preliminary

TTS Model Structure



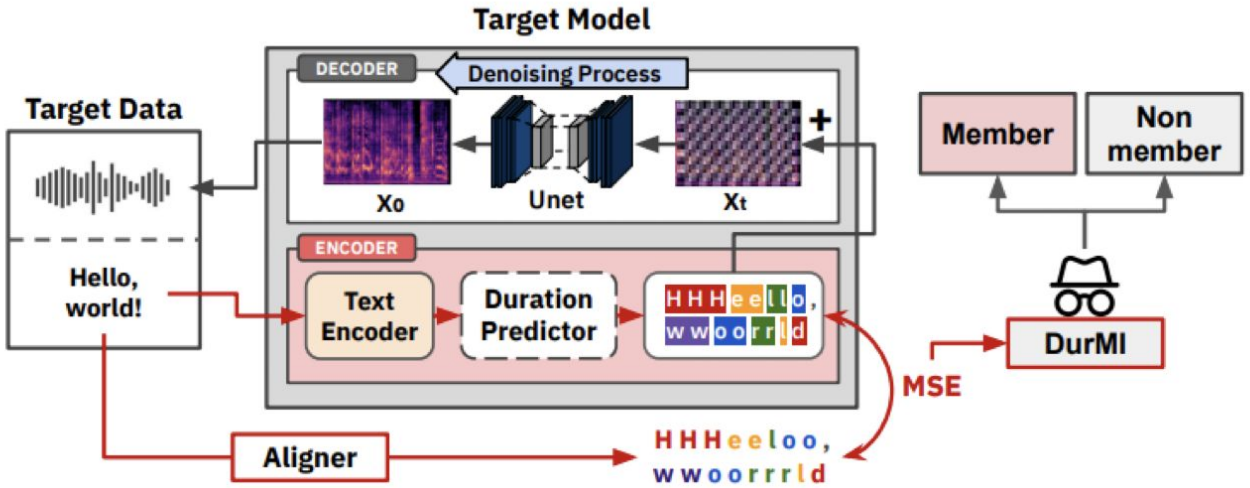
We target non-autoregressive TTS model with **alignment module**

EWHA, THE FUTURE WE CREATE

Methodology - Duration loss

Duration predictor

NAR TTS models have Duration predictor module to learn the alignment of phones.
In our work, we utilized duration loss from this module to conduct MIA



Experiment - Evaluation

- Performance of MIA methods on **GradTTS** across various datasets.

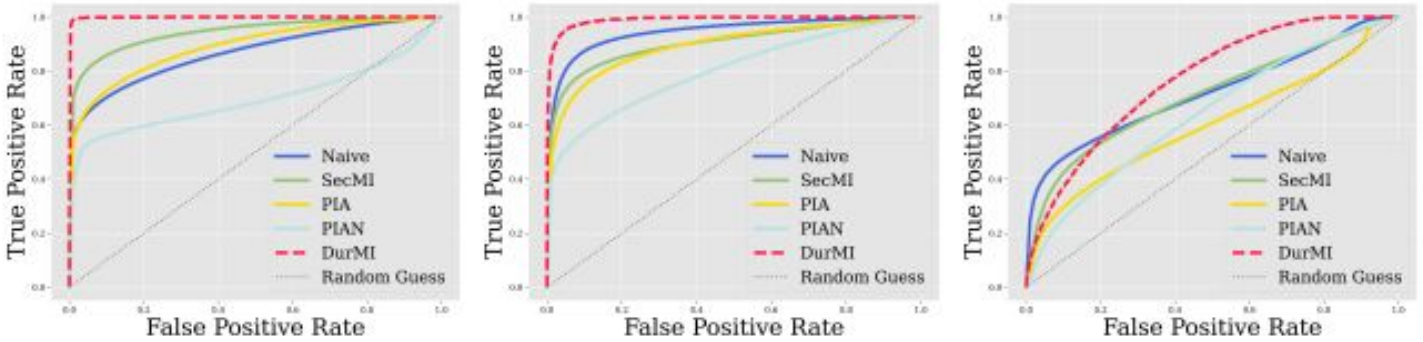
	LJSpeech		LibriTTS		VCTK	
	AUC	TPR@1% FPR	AUC	TPR@1% FPR	AUC	TPR@1% FPR
Naive Attack [15]	86.7	55.0	94.5	58.1	73.2	**29.5**
SecMI [16]	94.4	70.3	90.2	55.2	72.8	8.1
PIA [19]	89.0	55.0	89.3	47.0	64.4	9.7
PIAN [19]	69.0	37.4	81.8	37.4	66.6	6.1
DurMI	**99.8**	**98.9**	**98.9**	**83.5**	**76.8**	9.6

- Performance of MIA methods on **WaveGrad2** across various datasets.

	LJSpeech		LibriTTS		VCTK	
	AUC	TPR@1% FPR	AUC	TPR@1% FPR	AUC	TPR@1% FPR
Naive Attack	50.1	1.0	54.3	0.6	59.9	1.5
SecMI	49.4	1.0	47.6	0.3	55.4	1.0
PIA	50.8	0.4	51.7	0.1	52.1	0.8
PIAN	50.3	0.1	50.2	0.1	44.7	0.1
DurMI	**99.9**	**100.0**	**100.0**	**100.0**	**97.4**	**50.9**

Experiment - In various dataset

- **ROC curves** comparing MIA methods on the GradTTS model across various datasets.
  - LJSpeech (one speaker), LibriTTS, VCTK (multi-speaker)
  - Naive, SecMI, PIA, PIAN, Random guess & Our method



(a) LJSpeech (b) LibriTTS (c) VCTK

Experiment - Running time

- **Running time (in milliseconds) for performing MIA on a single sample.**
  - DurMI requires only a **single forward pass** before the decoder stage, making it over **100× faster than SecMI and more than 50× faster than PIA.**

	Inference Time (ms)				
	Naive	SecMI	PIA	PIAN	DurMI
GradTTS	1.5364	3.0418	1.5333	1.5289	**0.0305**
WaveGrad2	1.8333	3.8435	1.9430	1.7947	**0.0448**

Conclusion

- We present **DurMI, a novel white-box membership inference attack** that leverages **duration loss** in diffusion-based TTS models.
- In contrast to prior approaches that depend on decoder-side diffusion losses, **DurMI exploits alignment supervision signals available before the decoder stage**, achieving both higher inference **accuracy** and significantly **lower computational cost**.
- Evaluated across GradTTS and WaveGrad2, DurMI consistently **outperforms existing methods**, even on waveform-based models where prior attacks fail.
- This demonstrates that duration loss encodes strong, sample-specific signals and constitutes a vulnerable component in TTS training pipelines.

Further: TTS Model and Differential Privacy(DP)

Applying DP to TTS model as a defence to MIA

Step	Information	Should Apply DP?
Text Encoder	phone representation	No
Duration Predictor	duration	Yes <- new!
Speaker Embedding	speak identity	Yes, and already do
Decoder	latent audio	Yes