DURMI: DURATION LOSS AS A MEMBERSHIP SIGNAL IN TTS MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-speech (TTS) models such as FastSpeech2, Grad-TTS, and VITS2 achieve state-of-the-art quality but risk memorizing and leaking sensitive training data. Existing membership inference attacks (MIAs) for diffusion-based TTS models typically rely on denoising errors, which are costly to compute and weak at capturing sample-specific memorization.

We introduce DurMI, the first membership inference attack that exploits duration loss, a core alignment signal in TTS models, as a discriminative indicator of membership. Duration loss captures the model's tendency to overfit alignment targets, whether derived from deterministic aligners such as MAS and MFA or from stochastic predictors as in VITS2. Leveraging this signal, DurMI enables accurate inference with a single forward pass, while remaining broadly applicable across diverse TTS architectures.

Experiments across diverse architectures, including diffusion (Grad-TTS, WaveGrad2), flow-matching (VoiceFlow), transformer (FastSpeech2), and stochastic-duration (VITS2), on three benchmarks show that DurMI consistently outperforms prior MIAs, including on waveform-level synthesis where existing attacks fail. These results highlight DurMI's effectiveness, efficiency, and broad applicability, underscoring the need for privacy-preserving training in modern TTS systems.

1 INTRODUCTION

Text-to-speech (TTS) models have advanced rapidly, producing natural and high-quality speech across diverse architectures, including diffusion-based systems (Grad-TTS (Popov et al., 2021), WaveGrad2 (Chen et al., 2021)), flow-matching approaches (VoiceFlow (Guo et al., 2024)), transformer-based models (FastSpeech2 (Ren et al., 2020)), and stochastic-alignment models (VITS2 (Kong et al., 2023b)). These models are trained on large-scale datasets that often contain sensitive or proprietary content, raising critical concerns about privacy leakage.

Such risks are especially acute in applications like voice assistants and medical TTS, where training data may reveal personal identity, health-related utterances, or location cues (Chen et al., 2023). An adversary able to infer whether a specific utterance was used for training—a scenario formalized as a *membership inference attack (MIA)*—can compromise user privacy and even expose organizations to legal risks, particularly when copyrighted materials such as audiobooks are memorized.

MIA has been extensively studied in computer vision (Chen et al., 2020; Carlini et al., 2022; Li et al., 2024b) and natural language processing (Shi et al., 2023; Mattern et al., 2023; Fu et al., 2024), with extensions to GANs, VAEs (Hayes et al., 2017; Hilprecht et al., 2019; Sui et al., 2023), and diffusion models (Matsumoto et al., 2023; Duan et al., 2023; Dubiński et al., 2024; Hu & Pang, 2023). These methods assume training samples yield lower loss than non-members. However, state-of-the-art attacks like SecMI (Duan et al., 2023) require computing denoising errors across all timesteps, which is computationally prohibitive for large-scale TTS.

To improve efficiency, Proximal Initialization Attack (PIA) (Kong et al., 2023a) reduces the number of denoising steps and extends MIA to mel-spectrogram and waveform-level TTS models. However, these approaches treat TTS largely as a generic generative model, overlooking architectural features that are central to speech synthesis. In particular, *alignment mechanisms and duration supervision*

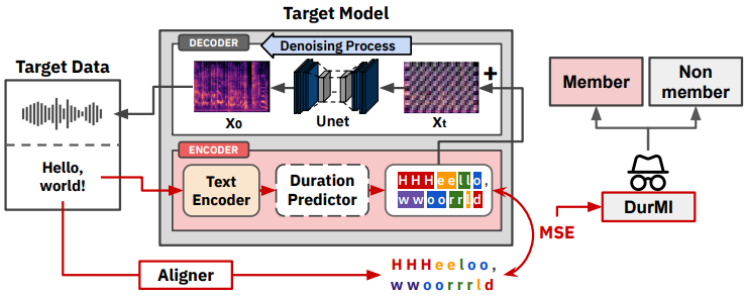


Figure 1: Overview of DurMI: illustrated here on a diffusion-based TTS model, where the difference between predicted and ground-truth durations from the aligner is used as a membership signal. DurMI, however, applies broadly across transformer-, flow-matching-, and stochastic-alignment models as well. It requires only a single forward pass up to the decoder stage (red arrow).

are unique to TTS pipelines and directly influence how models memorize training utterances, yet remain unexploited by prior MIAs.

In this work, we propose *DurMI* (Duration Loss-Based Membership Inference), the first attack to exploit *duration loss* as a discriminative signal for membership inference. Our key insight is that duration predictors are trained to match sample-specific alignment targets, whether deterministic (e.g., MAS (Kim et al., 2020), MFA (McAuliffe et al., 2017)) or stochastic as in VITS2 (Kong et al., 2023b), which encourages overfitting to utterance-level timing patterns. This makes duration loss a highly effective signal for membership inference. As shown in Figure 1, DurMI requires only a single forward pass before the decoder, bypassing diffusion entirely and yielding **50–100$\times$ speedups** over prior diffusion-based MIAs.

We adopt the white-box setting, the standard evaluation protocol in MIA (Duan et al., 2023; Kong et al., 2023a). This assumption reveals worst-case vulnerabilities critical for privacy auditing and defense design, and reflects realistic contexts such as internal audits, regulatory compliance, or fine-tuning on user data.

We evaluate DurMI on Grad-TTS, WaveGrad2, VoiceFlow, FastSpeech2, and VITS2 across three benchmark datasets (LJSpeech (Ito & Johnson, 2017), VCTK (Yamagishi et al., 2019), and LibriTTS (Zen et al., 2019)). Through these experiments, we establish DurMI's key contributions: (i) it consistently outperforms prior MIAs in terms of detection accuracy, (ii) it generalizes across diverse architectures, including VITS2 with stochastic alignment, (iii) it achieves high efficiency, requiring only a single forward pass (over 100$\times$ faster than SecMI), and (iv) it is modality-robust, applying equally well to spectrogram- and waveform-based synthesis.

Finally, while zero-shot and alignment-free TTS has emerged, recent studies show these models often degrade on complex text and reintroduce auxiliary alignment for intelligibility (Jiang et al., 2025; Neekhara et al., 2024). Together with ongoing deployment of duration-supervised systems in industry, this indicates that alignment remains central to current and near-future TTS. Although DurMI cannot be directly applied to alignment-free architectures, we discuss potential *proxy indicators* as a direction for future work.

By exposing privacy risks rooted in duration alignment, DurMI establishes a new attack surface and a foundation for defenses, advancing the understanding of vulnerabilities in modern TTS systems.

2 Related Work

Membership Inference Attacks. Membership inference attacks (MIAs) aim to determine whether a data sample was part of a model's training set. First studied in discriminative models via output confidence scores (Shokri et al., 2017), they have since been extended to generative settings. LOGAN (Hayes et al., 2017) showed leakage in GANs, and recent work in LLMs uses log-likelihoods of rare tokens as membership signals (Shi et al., 2023; Zeng et al., 2023; Carlini et al., 2021).

Table 1: Comparison of existing membership inference attacks targeting diffusion models.

Method	Key Idea	Sampling	Computational Cost
Naive Attack	Computes MSE between predicted and ground-truth noise	Stochastic (DDPM)	Moderate (1 query)
SecMI	Collects timestep-wise noise prediction errors for classification	Deterministic (DDIM)	High (full trajectory)
PIA	Reconstructs sample via DDIM, re-diffuses once to measure error	Deterministic (DDIM)	Moderate (1–2 queries)
PIAN	Normalized variant of PIA using L_1 norm	Deterministic (DDIM)	Moderate (1–2 queries)

For diffusion models, MIAs exploit reconstruction errors during denoising. The Naive Attack (Matsumoto et al., 2023) measures mean squared error between predicted and ground-truth noise. SecMI (Duan et al., 2023) aggregates per-timestep errors, improving accuracy under deterministic DDIM sampling but at high cost. PIA and PIAN (Kong et al., 2023a) reduce overhead by reconstructing samples with few steps, though with weaker robustness. Table 1 summarizes these methods.

These techniques are largely tailored to images and overlook signals unique to TTS. White-box analysis, although not always feasible in deployment, remains standard in MIA (Duan et al., 2023; Kong et al., 2023a) and is critical for revealing worst-case vulnerabilities that guide auditing and defenses.

Text-to-Speech Models. Early autoregressive models such as WaveNet (van den Oord et al., 2016) and Tacotron (Wang et al., 2017) achieved high fidelity but suffered from slow inference. Non-autoregressive (NAR) models like FastSpeech2 (Ren et al., 2020) introduced explicit duration predictors to align text and audio, enabling parallel and controllable synthesis. Duration-based supervision is now widely adopted in many NAR systems, and our work is the first to show that this alignment loss itself leaks membership information.

Grad-TTS (Popov et al., 2021) was the first diffusion-based TTS model to generate mel-spectrograms via latent-space denoising, while WaveGrad2 (Chen et al., 2021) directly synthesized raw audio waveforms without relying on intermediate spectrogram representations. Building on diffusion, flow-matching methods (Mehta et al., 2024; Chen et al., 2024) have recently been introduced to TTS, aiming to improve training stability and sampling efficiency. For instance, VoiceFlow (Guo et al., 2024) leverages flow matching to enable faster and more robust speech synthesis compared to traditional diffusion models. In addition, VITS2 (Kong et al., 2023b) enhances the VITS (Kim et al., 2021) framework by incorporating stochastic duration modeling, which enables more diverse prosodic patterns and improves the naturalness of generated speech.

Recently, alignment-free and zero-shot systems such as E2-TTS (Eskimez et al., 2024) and F5-TTS (Chen et al., 2024) have emerged. Yet their robustness is debated. MegaTTS 3 (Jiang et al., 2025) finds such models degrade on complex text and improves quality by reintroducing sparse alignment. Similarly, NVIDIA's T5-TTS (Neekhara et al., 2024) reduces hallucinations and omissions by enforcing monotonic alignment. These results suggest alignment will remain complementary, if not central, in future TTS. Although DurMI cannot be applied directly to zero-shot models, they still rely on implicit alignment cues such as discrepancies between target and generated utterance length. We highlight these as potential *proxy indicators* for extending MIA to implicit-alignment systems, and provide further discussion in Section 5.1 and Appendix A.7.

3 PRELIMINARIES

This section reviews the role of duration loss and alignment strategies in representative TTS architectures: Grad-TTS, WaveGrad2, VITS2, VoiceFlow, and FastSpeech2.

3.1 DURATION LOSS

A large class of modern TTS models generally consist of three modules: a text encoder, a duration predictor, and a decoder. While auxiliary objectives such as pitch or energy vary across systems, *duration loss is universally present*. It supervises phoneme-to-frame alignment, enforcing sample-specific timing patterns. DurMI builds directly on this shared mechanism: by comparing predicted and ground-truth durations, it extracts membership signals that generalize across diverse architectures.

Grad-TTS. Durations $d \in \mathbb{R}^L$ are obtained via MAS, which computes sample-specific phoneme-to-frame mappings. The model minimizes

$$\mathcal{L}_{\text{dur}}^{\text{GT}} = \left\| f_{\text{dur}}(\text{sg}[f_{\text{enc}}(c)]) - d \right\|_2, \tag{1}$$

where $\text{sg}[\cdot]$ blocks gradients to the encoder. Because MAS adapts dynamically during training, it introduces variability across utterances, slightly weakening membership leakage but improving synthesis quality.

WaveGrad2. Here, durations d are precomputed using MFA, a fixed alignment tool. The predictor minimizes

$$\mathcal{L}_{\text{dur}}^{\text{WG}} = \| \log \hat{d} - \log d \|_2. \tag{2}$$

Unlike MAS, MFA provides static, non-adaptive alignments, which tend to overfit to training utterances. This makes duration loss in WaveGrad2 a stronger membership signal.

VITS2. VITS2 combines MAS-based targets with adversarial learning. Predicted durations $\hat{d}$ are optimized as

$$\mathcal{L}_{\text{dur}}^{\text{V2}} = \text{MSE}(\hat{d}, d) + \lambda L_{\text{adv}}(G), \tag{3}$$

where L_{adv} encourages natural duration distributions. This stochastic training setup reduces determinism but still preserves sample-specific information, showing that DurMI generalizes beyond purely deterministic predictors.

VoiceFlow and FastSpeech2. Both rely on forced alignments (e.g., MFA) to generate ground-truth durations and minimize a mean squared error:

$$\mathcal{L}_{\text{dur}}^{\text{VF,FS2}} = \frac{1}{N} \sum_{i=1}^{N} \|\hat{d}_i - d_i\|^2. \tag{4}$$

These models provide stable and explicit supervision, making duration loss a reliable membership signal.

3.2 ALIGNMENT MECHANISMS FOR DURATION PREDICTION

Montreal Forced Aligner (MFA). MFA is an offline, non-differentiable aligner based on Gaussian Mixture Model–Hidden Markov Model (GMM-HMM) acoustic models and MFCC features, implemented in Kaldi. It produces fixed phoneme-to-frame alignments independent of model parameters and is widely used in TTS pipelines. These static alignments often amplify memorization signals.

Monotonic Alignment Search (MAS). MAS is a differentiable dynamic programming algorithm that finds monotonic text–audio alignments by maximizing cumulative likelihood:

$$Q_{i,j} = \max(Q_{i-1,j-1}, Q_{i,j-1}) + \log \mathcal{N}(z_j; \mu_i, \sigma_i),$$

where z_j is an acoustic frame and (μ_i, σ_i) are phoneme-level Gaussian parameters. Backtracking recovers the alignment path A^*, and durations are computed as

$$d_i = \log \left(\sum_{j=1}^{F} \mathbb{I}\{A^*(j) = i\} \right).$$

Unlike MFA, MAS integrates alignment into training, yielding more adaptive but less deterministic targets, which weakens overfitting signals.

Stochastic Duration Modeling. VITS2 introduces stochastic duration prediction to capture natural variability in rhythm and prosody. Durations are generated as

$$\hat{d} = G(z_d, h_{\text{text}}), \quad z_d \sim \mathcal{N}(0, I),$$

where G is trained with a combination of mean squared error and adversarial loss:

$$\mathcal{L}_{\text{dur}}^{\text{V2}} = \text{MSE}(\hat{d}, d) + \lambda L_{\text{adv}}(G).$$

This formulation reduces determinism but still preserves sample-specific timing, showing that alignment supervision remains embedded even in stochastic predictors.

4 Duration Loss-Based Membership Inference

We introduce DurMI, a white-box membership inference attack that leverages *duration loss* as a discriminative signal in TTS models. Our key insight is that duration predictors are trained to minimize sample-specific alignment errors – leading to potential overfitting – which can be exploited for identifying training membership.

4.1 Threat Model and Assumptions

We assume a white-box threat model in which the adversary has full access to a trained TTS model, including the encoder f_{enc}, duration predictor $f_{\text{dur}}(\cdot;\theta)$, and the loss function $\mathcal{L}_{\text{dur}}$. Given a target sample $x = (c, a)$ and its alignment target d (obtained from MAS or MFA), the goal is to determine whether x belongs to the training set $\mathcal{D}_{\text{train}}$.

4.2 Formulation of DurMI

Let $f_{\text{enc}}(c)$ denote the phoneme-level representation of the input text sequence c, and let $f_{\text{dur}}(\cdot;\theta)$ be the duration predictor parameterized by θ. The duration loss for input x is computed as:

$$\mathcal{L}_{\text{dur}}(x;\theta) = \left\| f_{\text{dur}}(\text{sg}[f_{\text{enc}}(c)];\theta) - d \right\|_p, \tag{5}$$

where d is the ground-truth log-duration vector and $\text{sg}[\cdot]$ is the stop-gradient operator that blocks gradients during optimization. The norm $\|\cdot\|_p$ is selected by the attacker (typically $p = 2$).

The adversary then defines a binary membership function $\mathcal{M}$ based on thresholding the loss:

$$\mathcal{M}(x) = \begin{cases} 1 & \text{if } \mathcal{L}_{\text{dur}}(x;\theta) < T \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where T is a decision threshold estimated via a calibration set or a shadow model. Because duration predictors are often overfitted to training samples – especially when supervised with deterministic alignment targets – samples with lower loss are more likely to be members.

4.3 Comparison with Diffusion-Based MIAs

Table 2 compares DurMI against existing MIA techniques that rely on diffusion loss or timestep-wise noise prediction errors. While these prior approaches are applicable to generic diffusion models, they often suffer from high computational cost and require fine-grained calibration. In contrast, DurMI offers a TTS-specific yet efficient and highly discriminative alternative.

4.4 Advantages of Duration Loss for Membership Inference

While prior MIAs on diffusion models (Matsumoto et al., 2023; Duan et al., 2023; Kong et al., 2023a) focus on reconstruction loss or noise prediction errors, duration loss offers several key advantages in the TTS setting.

Table 2: Comparison of duration loss (DurMI) and diffusion-based membership signals.

Aspect	DurMI	Diffusion-based MIAs
Target signal	Duration loss	Noise prediction error
Sample specificity	High	Low to moderate
Loss variance	Low	High
Computational cost	Low (single forward pass)	High (multi-step rollouts)

Table 3: Comparison of intra-class variance, inter-class variance, and LDA scores for duration loss (DurMI) and diffusion loss (Naive Attack, PIA) used as membership signals.

Method	Intra-class variance	Inter-class variance	LDA
DurMI	**0.002**	0.012	**6.0**
Naive Attack	6.658	0.031	0.004
PIA	27.549	0.203	0.007

First, duration loss is sample-specific, whether derived from deterministic aligners (e.g., MAS, MFA) or stochastic predictors (e.g., VITS2). Deterministic alignments encourage exact overfitting to utterance-level timing, while stochastic predictors still rely on training-conditioned distributions that retain sample-level bias. In both cases, the duration loss tightly encodes alignment signals tied to individual training examples, unlike diffusion loss, which is distributional and exhibits higher intra-class variance.

Second, duration loss exhibits stronger separability. As shown in Table 3, duration loss has significantly lower intra-class variance and higher inter-class separability, as quantified by Fisher's Linear Discriminant Analysis (LDA) score – a metric that captures how well two classes (member vs. non-member) are separated based on the ratio of between-class to within-class variance. This enhanced separation facilitates simple and effective threshold-based inference.

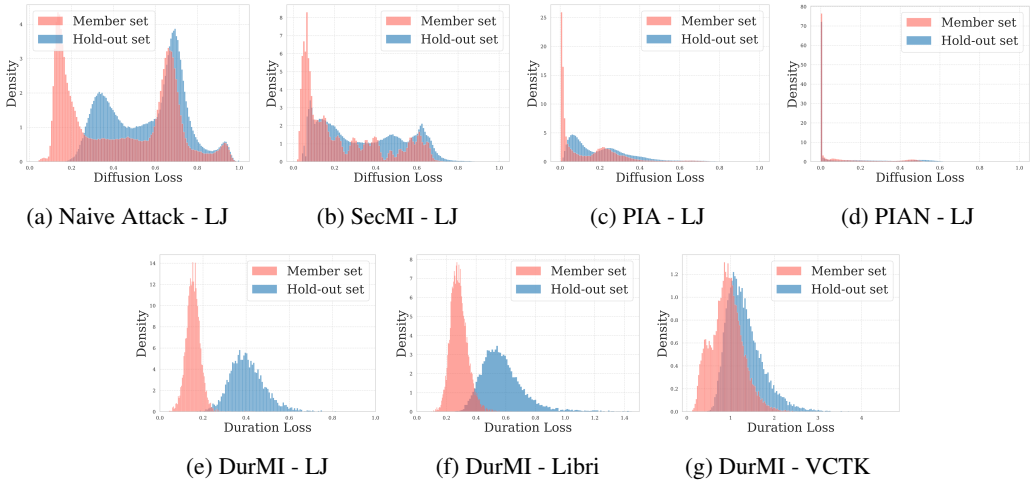
Figure 2 further visualizes the distributional separation between member and non-member samples across various MIA techniques. DurMI produces sharper decision margins for LJSpeech and LibriTTS, while diffusion-based losses exhibit substantial overlap, particularly for LJSpeech. DurMI shows greater overlap on the VCTK dataset, which is discussed in more detail in Appendix A.6.

Finally, DurMI is computationally efficient. Unlike prior attacks that require full diffusion rollouts or multiple inference passes (e.g., SecMI, PIA), DurMI computes a single forward pass prior to the decoder, independent of decoder-level sampling. This makes DurMI particularly practical for large-scale TTS systems.

5 Experiments

We evaluate DurMI on five representative TTS architectures: two diffusion-based models (Grad-TTS and WaveGrad2), one transformer-based model (FastSpeech2), one flow-matching model (VoiceFlow), and one stochastic-duration model (VITS2). Experiments are conducted on three widely used benchmarks: LJSpeech (Ito & Johnson, 2017), VCTK (Yamagishi et al., 2019), and LibriTTS (Zen et al., 2019). Each experiment is repeated three times, and averages are reported, as standard deviations are consistently below 0.1%. Baseline comparisons include Naive Attack, SecMI, PIA, and PIAN, with full implementation details and preprocessing procedures provided in Appendix A.1.

To ensure fair evaluation, we split each dataset evenly into member samples (50%, used for training) and non-member samples (50%, held out). From both pools, 20% are further used as a calibration set and 80% as an evaluation set. The calibration set assumes that the adversary has access to a small subset of both member and non-member data, as is standard in MIA research, and is used solely to set decision thresholds. Evaluation samples are never used for calibration, ensuring that AUROC and TPR@1% FPR reflect unbiased attack performance.



(a) Naive Attack - LJ (b) SecMI - LJ (c) PIA - LJ (d) PIAN - LJ

(e) DurMI - LJ (f) DurMI - Libri (g) DurMI - VCTK

Figure 2: Member vs. Non-member distribution separability using diffusion loss (Naive, SecMI, PIA, PIAN) vs. duration loss (DurMI) across datasets: LJSpeech (LJ), LibriTTS (Libri), and VCTK.

Table 4: 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	86.7	55.0	94.5	58.1	73.2	**29.5**
SecMI	94.4	70.3	90.2	55.2	72.8	8.1
PIA	89.0	55.0	89.3	47.0	64.4	9.7
PIAN	69.0	37.4	81.8	37.4	66.6	6.1
DurMI	**99.7**	**99.1**	**98.9**	**82.8**	**86.7**	18.2

We report two standard metrics used in membership inference literature (Carlini et al., 2022; 2023; Li et al., 2024a). The first is AUC which measures the overall ability of the attack to distinguish members from non-members. The second is TPR@1% FPR, which quantifies the true positive rate under strict false positive constraints, emphasizing precision in privacy-sensitive settings.

Compared to Baselines. DurMI consistently outperforms baseline MIA methods across models and datasets, achieving the highest AUC and TPR@1%FPR as shown in Tables 4 and 6 with the exception of Grad-TTS on VCTK set, which we analyze in detail later. ROC curves in Figure 4 confirm this trend, with DurMI maintaining superior detection rates across a wide range of false positive rates. Importantly, DurMI is the only method that achieves strong performance on WaveGrad2, while existing methods perform near random guessing as shown in Table 5. This highlights the advantage of operating on alignment signals upstream of the decoder, rather than relying on modality-sensitive denoising losses. As shown in Table 6, DurMI achieves high AUC on VITS2, but its TPR@1% FPR is comparatively lower. This can be attributed to VITS2's stochastic duration modeling, which weakens attack precision. The effect is further amplified on multi-speaker datasets like LibriTTS and VCTK, where greater variability in speech patterns reduces detection accuracy at low FPR.

Efficiency Comparison. DurMI is significantly faster than all baseline methods. As shown in Table 7, it requires only a single forward pass through the duration predictor, bypassing the diffusion decoding process entirely. This makes DurMI approximately 100× faster than SecMI and 50× faster than PIA for per-sample inference.

Alignment Mechanisms. DurMI achieves better performance on WaveGrad2 than on Grad-TTS, which we attribute to differences in alignment. WaveGrad2 relies on MFA, a fixed aligner prone

Table 5: 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**	**47.0**

Table 6: Performance of DurMI on different TTS models across various datasets.

	LJSpeech		LibriTTS		VCTK	
Model	AUC	TPR@1% FPR	AUC	TPR@1% FPR	AUC	TPR@1% FPR
VoiceFlow	99.2	93.9	98.0	56.5	98.9	90.6
FastSpeech2	100.0	100.0	99.2	90.5	99.5	93.7
VITS2	97.5	80.1	85.5	22.4	87.1	12.2

to overfitting, thereby amplifying membership signals. In contrast, Grad-TTS employs MAS, a differentiable and adaptive aligner that reduces sample-specific memorization and weakens attack effectiveness. Notably, DurMI also performs strongly on VITS2, which adopts stochastic duration modeling rather than deterministic alignment, indicating that DurMI generalizes to probabilistic alignment strategies as well.

Dataset. On VCTK, DurMI attains high AUC but relatively lower TPR@1%FPR compared to other datasets. We attribute this to dataset-specific characteristics, including shorter utterances and lower text overlap between training and test sets. Full analysis of speaker composition, utterance length, and vocabulary overlap is provided in Appendix A.6.

5.1 Ablation Study

We conduct ablation studies on Grad-TTS using the VCTK dataset to examine the impact of various factors.

Training Epochs. As shown in Figure 2h, increasing the number of training epochs leads to overfitting and better MIA performance. The performance plateaus after 1,000 epochs, suggesting 1,000–2,000 epochs as a practical range.

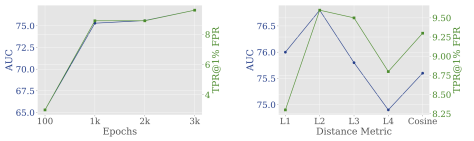
Distance Metric for Duration Loss. Figure 2i shows that L_2-norm (MSE) provides the best results. It aligns more closely with MAS-derived targets, strengthening memorization and member/non-member separation.

Sensitivity to Utterance Length. We grouped samples into two clusters based on the top and bottom 10% of utterance lengths and evaluated all four combinations of member and non-member clusters. DurMI consistently demonstrated clear separability across these settings, indicating robustness to input length. Detailed visualizations of separability across varying utterance lengths are provided in the Appendix A.5.

While DurMI directly exploits explicit duration predictors, recent advances in zero-shot and alignment-free TTS architectures raise the question of whether membership inference remains feasible without such modules. In these systems, the model typically receives text, audio context, and an additional input specifying the target generation length T_{gen}, then generates an output sequence of length T_{out}. This setup, illustrated in Figure 5a, mirrors the MIA scenario and motivates alternative attack signals.

Table 7: Running time (in milliseconds) for performing MIA on a single sample.

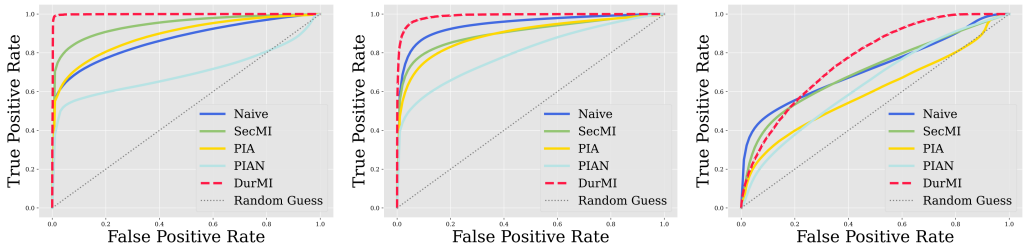
	Inference Time (ms)				
	Naive	SecMI	PIA	PIAN	DurMI
GradTTS	1.54	3.04	1.53	1.53	**0.03**
WaveGrad2	1.83	3.84	1.94	1.79	**0.04**



(h) Effect of training epochs.

(i) Effect of distance metrics.

Figure 3: Ablation study of DurMI.



(a) LJSpeech

(b) LibriTTS

(c) VCTK

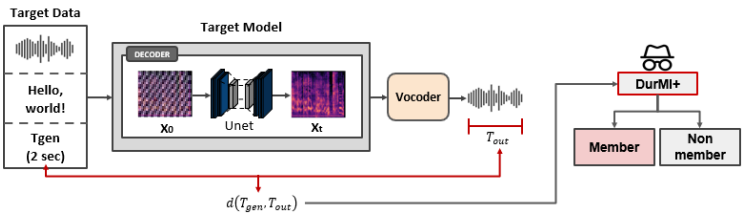
Figure 4: ROC curves comparing MIA methods on the Grad-TTS model across various datasets.

Even without explicit predictors, implicit alignment cues persist and can leak membership information. In particular, the discrepancy $d(T_{\text{gen}}, T_{\text{out}})$ is often smaller for training samples than for unseen inputs. Figure 5b highlights this distributional gap, suggesting $d(T_{\text{gen}}, T_{\text{out}})$ as a natural proxy indicator for membership inference.

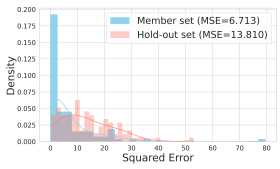
These observations indicate that alignment signals—whether explicit or implicit—are central to understanding privacy leakage in modern TTS. Although DurMI cannot be directly applied to zero-shot models, proxy indicators such as $d(T_{\text{gen}}, T_{\text{out}})$ provide a principled path to extend our methodology. In addition, this approach operates without requiring white-box access. Additional candidate indicators and detailed analysis are provided in Appendix A.7.

6 Conclusion

We present DurMI, a novel white-box membership inference attack that leverages duration loss in 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 Grad-TTS, WaveGrad2, FastSpeech2, VoiceFlow, and VITS2, DurMI consistently outperforms existing methods – including on waveform-based models where prior attacks fail – demonstrating that duration loss encodes strong sample-specific signals and represents a vulnerable component in TTS training pipelines.



(a) Overview



(b) Member vs. Non-member MSE distributions.

Figure 5: Extended DurMI: Overview of the attack and separation of member and non-member distributions based on MSE$(T_{\text{gen}}, T_{\text{out}})$

7 Ethics Statement

This work adheres to the ICLR Code of Ethics. No studies involving human subjects or sensitive data were conducted beyond standard publicly available datasets. Potential ethical considerations, including fairness, privacy, and research integrity, have been carefully evaluated, and no conflicts of interest or harmful outcomes are expected.

8 Reproducibility Statement

The complete codebase for our experiments is organized into three primary components: (1) `attack/`, containing model-specific MIA implementations; (2) `train/`, which includes training scripts for all TTS models; and (3) `README` files, offering step-by-step instructions for data preprocessing, model training, and evaluation.

We release all pretrained model checkpoints and the corresponding preprocessed datasets at `https://zenodo.org/records/15474571`. The release includes all model-dataset combinations (Grad-TTS, WaveGrad2, and VoiceFlow across LJSpeech, LibriTTS, and VCTK), totaling nine checkpoints. All MIA methods can be directly evaluated using the provided resources.

References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*, pp. 343–362, 2020.

Guangke Chen, Yedi Zhang, and Fu Song. Slmia-sr: Speaker-level membership inference attacks against speaker recognition systems. *arXiv preprint arXiv:2309.07983*, 2023.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660*, 2021.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *International Conference on Machine Learning*, pp. 8717–8730. PMLR, 2023.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemyslaw Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4860–4869, 2024.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 682–689. IEEE, 2024.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11121–11125. IEEE, 2024.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv preprint arXiv:1705.07663*, 2017.

Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019.

Hailong Hu and Jun Pang. Loss and likelihood based membership inference of diffusion models. In *International Conference on Information Security*, pp. 121–141. Springer, 2023.

Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Bai Jionghao, Xiaoda Yang, Jialong Zuo, et al. Megatts 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*, 2025.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv preprint arXiv:2305.18355*, 2023a.

Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. *arXiv preprint arXiv:2307.16430*, 2023b.

Jingwei Li, Jing Dong, Tianxing He, and Jingzhao Zhang. Towards black-box membership inference attack for diffusion models. *arXiv preprint arXiv:2405.20771*, 2024a.

Qiao Li, Xiaomeng Fu, Xi Wang, Jin Liu, Xingyu Gao, Jiao Dai, and Jizhong Han. Unveiling structural memorization: Structural membership inference attack for text-to-image diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10554–10562, 2024b.

Tomoya Matsumoto, Takayuki Miura, and Naoto Yanai. Membership inference attacks against diffusion models. In *2023 IEEE Security and Privacy Workshops (SPW)*, pp. 77–83. IEEE, 2023.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. *arXiv preprint arXiv:2305.18462*, 2023.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pp. 498–502, 2017.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11341–11345. IEEE, 2024.

Paarth Neekhara, Shehzeen Hussain, Subhankar Ghosh, Jason Li, Rafael Valle, Rohan Badlani, and Boris Ginsburg. Improving robustness of llm-based speech synthesis by learning monotonic alignment. *arXiv preprint arXiv:2406.17957*, 2024.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning*, pp. 8599–8608. PMLR, 2021.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Hao Sui, Xiaobing Sun, Jiale Zhang, Bing Chen, and Wenjuan Li. Multi-level membership inference attacks in federated learning based on active gan. *Neural Computing and Applications*, 35(23): 17013–17027, 2023.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016. URL https://arxiv.org/abs/1609.03499.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards end-to-end speech synthesis, 2017. URL https://arxiv.org/abs/1703.10135.

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. Exploring memorization in fine-tuned language models. *arXiv preprint arXiv:2310.06714*, 2023.

A Appendix

A.1 Experimental Setup

We trained two diffusion-based TTS models (Grad-TTS and WaveGrad2), one transformer-based model (FastSpeech2), one flow-matching-based model (VoiceFlow), and a stochastic duration model (VITS2) on three benchmark datasets: LJSpeech Ito & Johnson (2017), LibriTTS Zen et al. (2019), and VCTK Yamagishi et al. (2019). LJSpeech is a single-speaker dataset containing approximately 13,000 short utterances. In contrast, VCTK consists of recordings from 110 English speakers totaling around 43,000 samples, while a 20,000-sample subset of LibriTTS – a large-scale multi-speaker corpus – was used in our experiments.

Each dataset was split into two disjoint subsets: one for training (member set) and the other for evaluation (non-member set). All models were trained using a batch size of 16, a learning rate of $1 \cdot e^{-4}$, and diffusion time steps – 50 for Grad-TTS and 1000 for WaveGrad2. All experiments were conducted on a single NVIDIA RTX A6000 GPU (48 GB VRAM), using the original model hyperparameters described below.

Grad-TTS. Grad-TTS was trained using a 22.05 kHz sampling rate and 80-dimensional mel-spectrograms, with an Fast Fourier Transform (FFT) size of 1024 and a hop length of 256. The encoder architecture comprises six convolutional layers (kernel size of 3, 192 channels), followed by two-headed multi-head attention and a dropout rate of 0.1.

WaveGrad2. WaveGrad2 was trained using a 22.05 kHz sampling rate and a hop length of 300. Its encoder consists of three convolutional layers with a kernel size of 5, 512 channels, and a dropout rate of 0.5.

VoiceFlow. VoiceFlow was trained on the same paired text-audio datasets as Grad-TTS and WaveGrad2, using a 16 kHz sampling rate and 80-dimensional mel-spectrograms. The model was trained for 3,000 epochs with a batch size of 10 and a learning rate of $5 \cdot 10^{-5}$. Its encoder consists of six layers, each with 192 channels, a kernel size of 3, a dropout rate of 0.1, and two-headed multi-head attention. The filter channel size was set to 768, and the hop length to 200.

FastSpeech2. FastSpeech2 adopts a transformer-based architecture with four encoder layers and six decoder layers, each with a hidden size of 256 and two attention heads. The feed-forward network employs a filter size of 1024 with convolutional kernels of size [9, 1], while both encoder and decoder apply a dropout rate of 0.2. Variance predictors for pitch and energy use a filter size of 256, kernel size of 3, and a dropout rate of 0.5, with linear quantization into 256 bins.

VITS2. VITS2 was trained with a sampling rate of 22.05 kHz, using 80-dimensional mel-spectrograms (FFT size 2048, hop length 256, window length 1024). The model integrates variational inference with stochastic duration modeling, employing a transformer-based text encoder with six layers, two attention heads, and hidden dimensionality of 192. The decoder incorporates eight normalizing flows and a HiFi-GAN–style generator with multi-scale residual blocks (kernel sizes 3, 7, 11). Notably, unlike deterministic duration models, VITS2 uses stochastic duration prediction, enabling more diverse prosody modeling.

A.2 Baseline MIA Configurations

Grad-TTS and VoiceFlow are continuous-time diffusion models, where the diffusion timestep t is sampled from the interval $[0, 1]$. For all attack methods – Naive Attack, SecMI, and PIA – we fix the number of diffusion timesteps to 100. Following their original implementations, both the Naive Attack and SecMI compute sample-wise reconstruction errors using the ℓ_2 norm between the predicted and ground-truth noise at each timestep. In contrast, PIA adopts the ℓ_4 norm to place greater emphasis on large errors, thereby increasing sensitivity to outliers in the denoising process.

In contrast, WaveGrad2 is a discrete-time diffusion model. According to the original codebase, the Naive Attack is performed with 100 discrete timesteps. For SecMI and PIA, the diffusion process is run with 1,000 timesteps, from which 100 are uniformly sampled at intervals of 10 to reduce computational overhead. The same norm configurations are applied: ℓ_2 for Naive and SecMI, and ℓ_4 for PIA.

A.3 Text and Audio Data Preprocessing and Alignments

The preprocessing stage in TTS models involves processing both the encoder inputs (text) and decoder inputs (audio), as well as computing phoneme-to-audio alignments through an aligner to estimate target durations. Below, we describe the preprocessing and alignment procedures adopted for Grad-TTS, WaveGrad2, and VoiceFlow.

Grad-TTS Grad-TTS does not require explicit text normalization, and the audio can be used at its native sampling rate without resampling: 22,050 Hz for LJSpeech, 16,000 Hz for LibriTTS, and 48,000 Hz for VCTK. For alignment, Grad-TTS employs MAS to estimate target durations, implemented through the `monotonic_align` module compiled with Cython.

WaveGrad2 Text inputs are normalized by lowercasing and removing punctuation. All audio waveforms are resampled to a consistent sampling rate of 22,050 Hz. Phoneme-to-audio alignments are generated using the MFA, which outputs alignment data in the TextGrid format. A TextGrid

Table 8: Performance of DurMI across datasets and models.

Model	Dataset	AUC	TPR@1% FPR
GradTTS	LJSpeech	99.8 ± 0.0	98.83 ± 0.06
	LibriTTS	98.9 ± 0.0	83.5 ± 0.0
	VCTK	76.8 ± 0.0	9.6 ± 0.0
WaveGrad2	LJSpeech	99.9 ± 0.0	100.0 ± 0.0
	LibriTTS	100.0 ± 0.0	100.0 ± 0.0
	VCTK	97.4 ± 0.0	50.97 ± 0.06

file includes tiered time-aligned annotations (e.g., phoneme and word levels), specifying the start and end time of each phoneme within the audio. These alignments are used to extract precise phoneme durations for training. Pre-generated TextGrid files for all datasets are provided and can be accessed at the following link: `https://drive.google.com/drive/folders/10eUTzOU06gTRMiQPoyw-Yctflms3ZLTJ?usp=sharing`.

VoiceFlow To train the VoiceFlow model, the dataset must be organized in the Kaldi-style format. Accordingly, the preprocessing pipeline consists of two main stages: (1) metadata generation and (2) audio feature extraction. The following manifest files are created to structure and describe the dataset:

- `wav.scp`: It maps each utterance ID (typically the filename) to its corresponding audio file path.
- `utts.list`: It lists all utterance IDs extracted from `wav.scp`.
- `utt2spk`: It associates each utterance ID with a speaker ID. For single-speaker datasets like LJSpeech, the same speaker ID is used for all utterances.
- `text`: It contains pairs of utterance IDs and their corresponding transcript texts.
- `phn_duration`: It provides phoneme-level alignments, specifying the start time and duration of each phoneme within an utterance. These are extracted from TextGrid files generated by the MFA.

After metadata creation, mel-spectrogram features are extracted from the audio data using VoiceFlow's feature extractor. The features are stored in the following Kaldi-compatible formats:

- `feats.ark`: A binary file containing the actual mel-spectrogram feature matrices.
- `feats.scp`: A text file mapping each utterance ID to the corresponding entry in the `.ark` file.

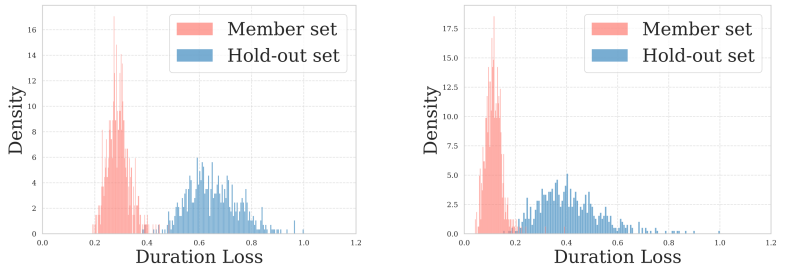
VoiceFlow supports phoneme alignment using either MAS or MFA. During training, it uses phoneme durations generated by MAS, rather than the precomputed durations from `phn_duration`. Finally, the phoneme-level transcripts are aligned to phone IDs listed in `phones.txt`, which are used as input to the model.

A.4 Statistical Significance Analysis

To ensure the reliability and robustness of our findings, we conducted each DurMI experiment three times across both Grad-TTS and WaveGrad2 models, using all three datasets: LJSpeech, LibriTTS, and VCTK. The aggregated results, including mean values and standard deviations, are presented in Table 8. All iterations produced consistent outcomes, with negligible standard deviations (less than 0.1%).

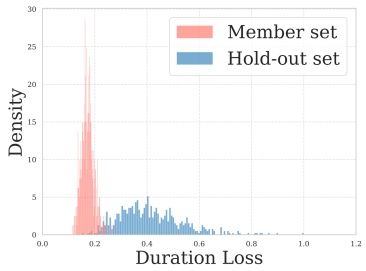
A.5 Ablation Study: Impact of Utterance Length

TTS models typically incorporate a phoneme-level duration predictor, trained to minimize the discrepancy between predicted and actual phoneme durations. Longer utterances, which contain more

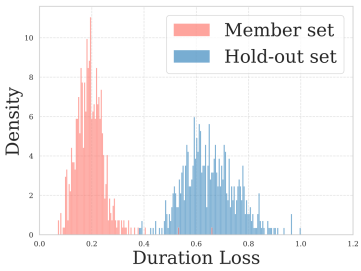


(a) Member set: Top 10%, Non-Member set: Top 10%

(b) Member set: Bottom 10%, Non-Member set: Bottom 10%



(c) Member set: Top 10%, Non-Member set: Bottom 10%



(d) Member set: Bottom 10%, Non-Member set: Top 10%

Figure 6: Performance comparison of DurMI across different utterance length clusters.

Table 9: Membership inference performance across different utterance length conditions. Each subset contains approximately 600 samples.

Condition	**AUC**	**TPR @1% FPR**
Full dataset	99.8	98.9
Cluster 1	99.9	98.9
Cluster 2	99.8	98.9
Cluster 3	99.4	94.9
Cluster 4	99.8	100.0

phonemes, are prone to cumulative prediction errors. In addition, variations in pronunciation and prosody increase the modeling complexity for longer utterances. If such utterances are underrepresented in the training data, the model's generalization ability to these cases may be further limited.

Based on these considerations, we hypothesize that utterance length could influence the success rate of membership inference attacks (MIA). To investigate this, we conducted an ablation analysis using the Grad-TTS model and the LJSpeech dataset. We divided the dataset into clusters based on utterance length, selecting the top and bottom 10% of utterances. We then evaluated MIA performance across the following four cluster combinations:

1. Both member and non-member samples belong to the top 10% of utterance lengths (Cluster 1).
2. Both member and non-member samples belong to the bottom 10% (Cluster 2).
3. Member samples are from the top 10%, and non-member samples from the bottom 10% (Cluster 3).
4. Member samples are from the bottom 10%, and non-member samples from the top 10% (Cluster 4).

Table 10: GradTTS-VCTK AUROC and FPR@1%TPR under different overlap conditions.

	Speaker Overlap Present	Speaker Overlap Absent
Text Overlap Present	96.8 / 9.6	82.25 / 10.47
Text Overlap Absent	85.90 / 22.04	85.91 / 22.27

Table 11: Speaker composition across datasets.

Dataset	Total Speakers	Member	Non-member	Overlap
LJSpeech	1	1	1	1
VCTK	50	25	26	1
LibriTTS	185	89	87	1

Figure 6 shows the distribution of duration loss across the four clusters. In all cases, member and non-member samples remain clearly separable.

Table 9 reports the AUC and TPR@1%FPR for each cluster. Although Cluster 3 shows a modest decrease (approximately 5 percentage points) in TPR@1%FPR, the overall impact of utterance length on DurMI performance is minimal.

In summary, the experimental results (Figure 6 and Table 9) indicate that utterance length does not significantly influence the effectiveness of MIA. Despite the reliance on duration information, the variation in utterance length yields only marginal changes in duration loss, suggesting limited impact on attack success.

A.6 Dataset Analysis and Overlap Effects

To contextualize the results reported in Tables 4 and 6, we provide additional analysis of dataset-specific factors, focusing on speaker/text overlap and corpus statistics. This supplementary analysis clarifies why certain datasets, particularly VCTK, exhibit relatively lower TPR@1%FPR despite maintaining high AUROC.

A.6.1 GradTTS on VCTK Under Overlap Conditions

Table 10 reports GradTTS performance on VCTK when speaker and text overlaps are selectively allowed or removed.

The results show three key trends: (1) performance is strongest when both text and speaker overlap, (2) removing both types of overlap produces the lowest TPR, and (3) absence of text overlap degrades detection more severely than absence of speaker overlap. This indicates that text overlap is a decisive factor for generalization and detection accuracy.

A.6.2 Dataset Composition

We also examine dataset-level statistics. Table 11 reports speaker splits, Table 12 shows average utterance length, and Table 13 summarizes vocabulary overlap.

- **Speakers:** VCTK includes 50 speakers with limited overlap, while LibriTTS has 185 speakers and LJSpeech only one.
- **Utterance length:** VCTK utterances average 3.2s, much shorter than LJSpeech (6.6s) or LibriTTS (12.7s), reducing temporal context.
- **Word overlap:** VCTK shows fewer shared words (4,694) compared to LJSpeech (8,631) and LibriTTS (12,633), limiting textual redundancy.

Table 12: Average utterance length (sec).

Dataset	Member	Non-member
LJSpeech	6.56	6.59
VCTK	3.24	3.30
LibriTTS	12.67	12.67

Table 13: Word overlap between member and non-member subsets.

Dataset	Member-only	Non-member-only	Shared
LJSpeech	7,123	7,189	8,631
VCTK	1,624	1,410	4,694
LibriTTS	8,221	7,789	12,633

A.6.3 Remarks

These analyses explain why VCTK yields lower TPR@1%FPR despite high AUROC. Short utterances and reduced word overlap limit both temporal and lexical cues, making membership detection harder. By contrast, datasets with greater redundancy (LJSpeech, LibriTTS) provide stronger signals. Overall, dataset composition—particularly text overlap—plays a central role in shaping membership inference performance.

A.7 Proxy Indicators for Implicit Duration Models

Recent non-autoregressive (NAR) zero-shot text-to-speech (TTS) models, including E2-TTS, F5-TTS, Seed-TTS, and MaskGCT, do not rely on explicit duration modules. Instead, they are trained using infilling objectives, where random audio segments are masked and reconstructed given the remaining context and full text. This procedure enables zero-shot synthesis across unseen text-speaker pairs, while implicitly learning alignment between text and audio sequences. Architecturally, models such as E2-TTS employ only a mel-spectrogram generator and a vocoder, without phoneme-level duration supervision.

Although DurMI cannot be directly applied to these systems, implicit alignment cues remain exploitable for membership inference. We identify two proxy indicators:

(1) Target generation length discrepancy. During inference, the model receives a target length T_{gen} and generates an output of length T_{out}. Training samples tend to yield closer matches, whereas unseen data often show larger deviations. A variety of distance functions can be used to quantify the discrepancy between the target and generated lengths:

$$d(N_{\text{gen}}, N_{\text{out}}) \in \left\{ |N_{\text{gen}} - N_{\text{out}}|,\ (N_{\text{gen}} - N_{\text{out}})^2,\ \frac{|N_{\text{gen}} - N_{\text{out}}|}{N_{\text{gen}}},\ \text{Huber}_\delta(N_{\text{gen}} - N_{\text{out}}),\ \text{KL}\big(P_{N_{\text{gen}}} \,\|\, P_{N_{\text{out}}}\big) \right\}.$$

Different distance functions capture different aspects of mismatch: absolute error measures raw deviation, squared error penalizes larger discrepancies more heavily, and the normalized ratio accounts for variability across utterance lengths. This flexibility allows the indicator to adapt to different evaluation needs and dataset characteristics.

(2) Masked segment reconstruction quality. Infilling-based training requires reconstructing masked frames. For training data, reconstruction tends to be more accurate, preserving pronunciation, timing, and speaker traits. Non-member samples often exhibit lower fidelity. Thus, similarity between original and reconstructed segments provides another membership signal.

Summary. Both indicators reflect how well the model has internalized specific training utterances: $d(T_{\text{gen}}, T_{\text{out}})$ captures implicit duration alignment, while reconstruction fidelity reflects learned acoustic detail. These serve as natural extensions of DurMI to implicit-alignment TTS. We leave systematic evaluation of these proxy indicators as an important direction for future work.