TTSlow: Slow Down Text-to-Speech with Efficiency Robustness Evaluations

Xiaoxue Gao, *Member, IEEE,* Yiming Chen, Xianghu Yue, *Member, IEEE,* Yu Tsao, *Senior Member, IEEE,* Nancy F. Chen, *Senior Member, IEEE,*

***Abstract*—Text-to-speech (TTS) has been extensively studied for generating high-quality speech with textual inputs, playing a crucial role in various real-time applications. For real-world deployment, ensuring stable and timely generation in TTS models against minor input perturbations is of paramount importance. Therefore, evaluating the robustness of TTS models against such perturbations, commonly known as adversarial attacks, is highly desirable. In this paper, we propose TTSlow, a novel adversarial approach specifically tailored to slow down the speech generation process in TTS systems. To induce long TTS waiting time, we design novel efficiency-oriented adversarial loss to encourage endless generation process. TTSlow encompasses two attack strategies targeting both text inputs and speaker embedding. Specifically, we propose TTSlow-text, which utilizes a combination of homoglyphs-based and swap-based perturbations, along with TTSlow-spk, which employs a gradient optimization attack approach for speaker embedding. TTSlow serves as the first attack approach targeting a wide range of TTS models, including autoregressive and non-autoregressive TTS ones, thereby advancing exploration in audio security. Extensive experiments are conducted to evaluate the inference efficiency of TTS models, and in-depth analysis of generated speech intelligibility is performed using Gemini. The results demonstrate that TTSlow can effectively slow down two TTS models across three publicly available datasets.**

***Index Terms*—Text-to-speech; Inference efficiency; Model robustness; Adversarial attack; Audio security.**

I. Introduction

TEXT-to-speech (TTS) is a task that aims to generate spoken speech based on a given text input [1]–[4]. TTS has achieved remarkable advancements from the autoregressive TTS models [5]–[9] to the non-autoregressive TTS approaches [10]–[19], neural vocoders [20]–[24] and the integration of diffusion models [25]–[29] as well as the speech style modeling [30]–[33] in recent years. These developments have led to diverse applications in virtual assistants, audiobooks, emotion understanding, voice-over narration, and navigation systems [34]–[38].

Real-time inference capability is crucial for a TTS system to be deemed production-quality [39], [40], as without it, the system becomes impractical for most TTS applications [35]. Efficiency robustness and audio security are critical components of various practical applications of TTS systems [41], [42]. Notably, excessively long speech generation in TTS systems can even significantly compromise security in various domains. In banking, delayed speech synthesis may hinder transaction verification and security notifications, which give fraudsters more time to exploit vulnerabilities, posing a heightened risk to bank security and customer assets [43], [44]. In home security, slow TTS can delay critical alerts (e.g., intruder detection, fire alarms), reducing system effectiveness and increasing the risk of harm and property damage [42]. In car security, prolonged speech generation can delay navigation instructions and critical alerts, posing safety hazards [45], [46]. Thus, TTS models must prioritize high-efficiency robustness to ensure usability, practicality, and effective security measures. However, current efficient TTS works mainly focus on improving inference speed with normal inputs [11], [12], [19], while the robustness of TTS models against minor perturbations remains largely unexplored.

Adversarial attack serves as a common and effective practice to evaluate the robustness of neural models recently for real-world applications [41], [42]. Adversarial attacks aim to elicit incorrect predictions through slightly altering the input data, thereby enabling the automatic detection of the flaws in existing neural models and revealing their vulnerabilities [47]–[50]. This attack process, in turn, aids in assessing and enhances neural model robustness [41], [51]–[53]. Motivated by the success of adversarial attacks and the under-studied robustness issue of TTS models, we propose TTSlow, a simple yet effective unified adversarial approach, to examine whether existing TTS models can provide stable and timely responses when confronted with maliciously perturbed inputs, termed as efficiency robustness.

Based on the observations that longer speech outputs lead to more inference steps and excessively long generation time in TTS models, we design TTSlow to automatically discover malicious inputs that can elicit endless speech generation process through nearly imperceptible input perturbations. Specifically, we propose a novel approach method to slow down text-to-speech generation, that we refer to as TTSlow, which consists of two novel attack techniques: TTSlow-text, which incorporates both character-swap attack and homoglyphs replacement attack, and TTSlow-spk, achieved through a speaker-oriented projected gradient descent attack. Extensive experiments on both autoregressive and non-autoregressive TTS models on three datasets show that both TTSlow-text and TTSlow-spk

Xiaoxue Gao and Nancy F. Chen are with Institute for Infocomm Research, Agency for Science, Technology, and Research (A*STAR), Singapore 138632 (e-mails: Gao_Xiaoxue@i2r.a-star.edu.sg and nfychen@i2r.a-star.edu.sg).

Yiming Chen and Xianghu Yue are with the Department of Electrical and Computer Engineering, National University of Singapore, National University of Singapore, Singapore 117583 (e-mails: yiming.chen@u.nus.edu and xianghu.yue@u.nus.edu). (*Corresponding author: Xianghu Yue*).

Yu Tsao is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan, and is also with the Department of Electrical Engineering, Chung Yuan Christian University, Taiwan. (e-mail: yu.tsao@citi.sinica.edu.tw).

can significantly harm the inference efficiency with longer speech, longer inference time, higher inference energy and an average attack success rate (ASR) of 90.39%.

The contributions of this paper include:

- **New Problem Characterization**: We identify a novel audio security problem within the TTS domain, focusing on the automatic detection of model flaws to advance the field towards more trustworthy TTS systems. To the best of our knowledge, this is the first study to systematically investigate into the robustness of inference efficiency in TTS systems.
- **Novel approaches**: We propose novel optimization-oriented loss objectives for both text and speaker embedding-based adversarial attacks. Our TTSlow approach is also designed to encompass both autoregressive and non-autoregressive scenarios in TTS models, offering distinct objective designs tailored to each scenario. This work represents the first successful implementation of near-human imperceptible adversarial attacks on TTS systems.
- **Comprehensive Experimentation**: We conduct a systematic evaluation of two proposed attack strategies on two TTS models across three publicly available datasets. Our findings underscore the need for future research aimed at enhancing and safeguarding the inference efficiency robustness of TTS models.

The rest of this paper is organized as follows. Section II describes related work on TTS and adversarial attacks, providing context for the proposed model design. Section III presents an overview of our proposed TTSlow approach and two proposed attack strategies. Section IV introduces the proposed objective functions for TTSlow in both autoregressive and non-autoregressive scenarios. Section V details the database and experimental setup. Section VI discusses the experiment results. Finally, Section VII concludes the study.

II. Related Work

We review text-to-speech models and adversarial attacks to establish the foundation of this work.

A. Text-to-speech Models

Text-to-speech (TTS) has garnered significant attention in recent times, powering numerous real-world applications [54]–[57]. In our investigation of the factors affecting TTS models' inference speed in practical applications, we delve into the length prediction mechanisms during the speech generation process. The determination of when a TTS model should stop generating speech distinguishes TTS models into two primary categories: autoregressive and non-autoregressive models [10]–[12]. Autoregressive TTS models typically stop speech generation based on the prediction of a stop token [5]–[9] , whereas non-autoregressive TTS models determine the length of the generated speech using a duration predictor [10]–[17].

Among these models, SpeechT5 stands out as a powerful autoregressive model, structured on an encoder-decoder architecture [58]. It demonstrates exceptional speech quality through extensive pre-training on large-scale unlabeled speech and text data, showcasing its potential for real-world applications. Conversely, VITS has gained fame and widespread adoption as a non-autoregressive model, utilizing variational inference with adversarial learning in an end-to-end TTS system [14], [15]. These models, SpeechT5 and VITS, are selected as representative backbone TTS models for this study due to their distinct approaches and strong performance in their respective categories.

B. Adversarial Attacks

Recently, the adversarial attack methods have been developed to evaluate the efficiency robustness of the machine learning models for real-world applications, and it has been intensively studied in computer vision [51], [52], machine translation [59], natural language processing [47]–[50], automatic speech recognition [41], [53], [60]–[63] and speaker identification [64], [65] domains.

Adversarial attacks can be classified into accuracy-oriented and efficiency-oriented attacks [47]–[52], [59]. Accuracy-based attacks target reducing the robustness of models by decreasing recognition performance, as seen in adversarial attacks on speech recognition models [41], [53], [60]–[63]. Efficiency-oriented attacks, like SlothSpeech [41], focus on diminishing system inference efficiency, significantly impacting speech recognition model performance. However, there is a notable lack of research exploring the resilience of TTS models against security threats (adversarial attacks) from both accuracy and efficiency perspectives. This paper aims to bridge the existing gap by proposing an adversarial attack method to analyze efficiency robustness for TTS systems.

III. TTSlow

In this section, we formulate the research problem and describe our proposed TTSlow approach, which includes two attack strategies: TTSlow-spk and TTSlow-text.

A. Problem Formulation

TTSlow aims to accomplish two objectives: (i) significantly increasing the computational time and reducing the inference efficiency for the victim TTS, where 'victim' refers to the TTS model being targeted for attack, and (ii) maintaining minimal perturbations in the generated output. With these goals in mind, we approach the problem as one of constrained optimization problem:

$$\Delta = \underset{\delta}{argmax}\, len_f(x+\delta) \quad s.t. ||\delta|| \leq \epsilon, \tag{1}$$

where x represents the given benign input, and ϵ denotes the maximum allowed adversarial perturbation. f is the victim TTS model and $len_f(\cdot)$ is the output sequence length of the victim TTS model. Our proposed approach, TTSlow, seeks to find the optimal perturbation Δ that reduces the efficiency while ensuring that the perturbation remains within the permissible threshold ϵ (i.e., nearly human unnoticeable).

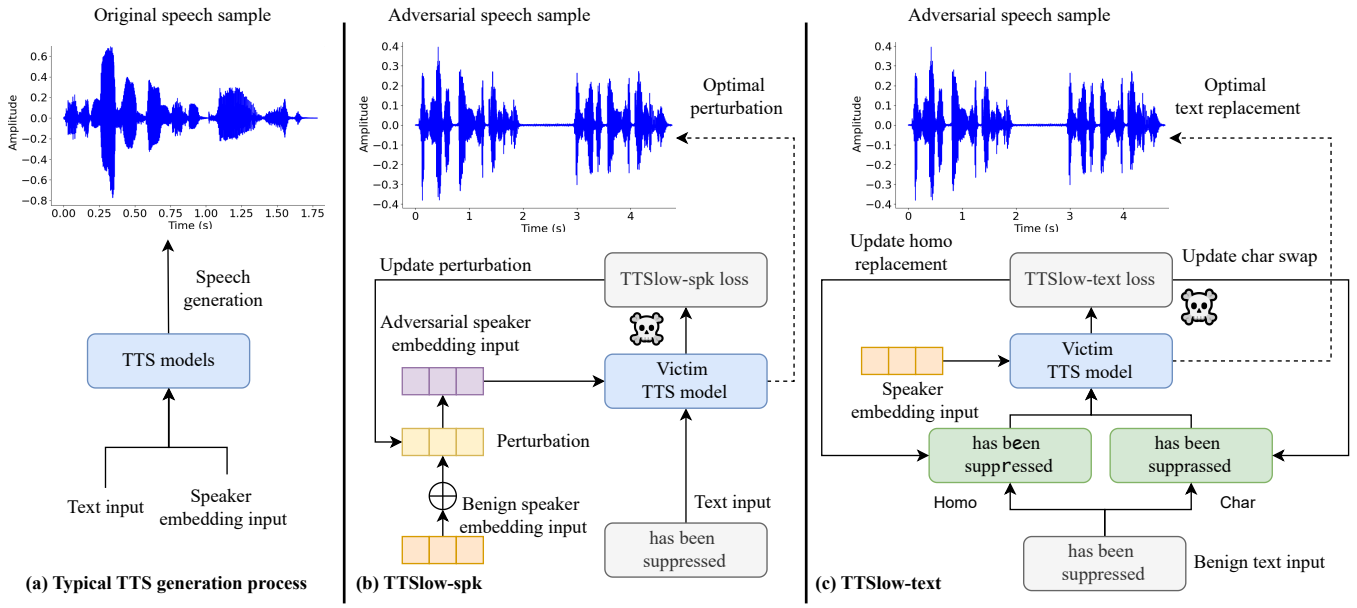


Fig. 1: The overview network architecture of (a) typical TTS generation process and the proposed TTSlow with two attack approaches: (b) TTSlow-spk and (c) TTSlow-text.

B. Overview of TTSlow

Fig. 1 shows an overview of proposed TTSlow against a typical multi-speaker TTS model. Conditioned on both benign text input and speaker embedding, the trained TTS models successfully synthesize approximately 1.75 seconds of synthesized speech, as depicted in Fig. 1 (a). Yet, when small perturbation is added to either the benign speaker embedding (TTSlow-spk in Fig. 1 (b)) or the benign text input (TTSlow-text in Fig. 1 (c)), the generated speech length suddenly increases to approximately 4.5 seconds, leading to longer generation process and a decrease in inference efficiency. The TTSlow approach employs two attack techniques to modify the given inputs and craft adversarial examples in Fig 1. The proposed attack techniques are tailored for various TTS scenarios: TTSlow-spk suits multi-speaker models with speaker embedding, TTSlow-text is applicable to both multi-speaker and single-speaker models using characters as inputs. These attack techniques will be further detailed in the subsequent sections.

C. TTSlow-spk

We begin by introducing the TTSlow-spk to assess the inference efficiency robustness of victim TTS models, as depicted in Fig. 1 (b). TTSlow-spk aims to generate an adversarial speaker embedding sample s_{adv} using projected gradient descent [66], [67] as the optimization approach while keeping the input text fixed.

1) Gradient Optimization: To slow down the speech generation, we propose to make the victim TTS model generate longer adversarial speech through updating the perturbation and optimizing the TTSlow-spk loss, denoted as $\mathcal{L}_{\text{TTSlow-spk}}$, aiming for no-stop speech generation. The TTSlow-spk attack is executed through iterating perturbations using gradient optimization with respect to the loss function $\mathcal{L}_{\text{TTSlow-spk}}$. The updated perturbation δ for each iteration is computed as follows:

$$\begin{aligned} s_{adv} &= s + \delta, \\ \delta &\leftarrow \Pi\left\{\delta - \alpha \cdot sign(\bigtriangledown_{\delta}\mathcal{L}_{\text{TTSlow-spk}}(f(s+\delta)))\right\} \end{aligned} \tag{2}$$

where s denotes the input speaker embedding, and α is the learning rate. $\bigtriangledown_{\delta}\mathcal{L}_{\text{TTSlow-spk}}$ is the gradient of the TTSlow-spk loss with respect to the perturbation δ. $\Pi\{\cdot\}$ is the projection function that enforces the ℓ_p constraint on the perturbation:

$$\Pi_{\ell_p}(s_{\text{adv}}) = \arg\min_{z \in S} \|s_{\text{adv}} - z\|_p, \tag{3}$$

where z is an element within the feasible set S, and $\|\cdot\|_p$ denotes the ℓ_p norm. $\Pi\{\cdot\}$ maps s_{adv} to the closest point z in S under the ℓ_p norm. We consider both ℓ_2 and ℓ_{inf} for distance norm.

2) Attack Methodology: After explaining the gradient optimization attack mathematically, Fig. 1 (b) visually depicts the process of adding perturbation to the benign speaker embedding input, leading to the creation of adversarial speaker embedding s_{adv}. This adversarial embedding is then fed into the victim TTS model to compute the TTSlow-spk loss. Further details regarding the TTSlow-spk loss will be provided in Section IV. After multiple iterations, the TTSlow-spk seeks to identify the optimal perturbation δ that minimizes the loss $\mathcal{L}_{\text{TTSlow-spk}}$ (thereby reducing the efficiency) while adhering to the perturbation constraint to remain the allowed threshold. This optimization process also aims to achieve near imperceptibility to humans, as measured by the distance norm.

D. TTSlow-text

Inspired by novel text replacement approaches in NLP [49], [68], we propose TTSlow-text, as illustrated in Fig. 1 (c), which combines two text-oriented attacks, homoglyphs replacement attack, and character swap attack. Homoglyphs are

unique characters that render the same glyph or a visually similar glyph. TTSlow-text is designed to maintain the original text length while generating adversarial samples that are almost imperceptible to humans.

1) Attack Methodology: Text-oriented attacks specifically target text inputs while keeping the speaker embedding input unchanged. TTSlow-text is proposed to iteratively modify the given text inputs to create adversarial examples in two strategies towards the TTSlow-text loss $\mathcal{L}_{\text{TTSlow-text}}$, as depicted in Fig. 1 (c). These two strategies include the character swap (char) attack, which replaces one character in the input text with another randomly selected character from the original character vocabulary, and the homoglyphs replacement (homo) attack, which replaces a character with its corresponding homoglyph. The homoglyph character mapping used in TTSlow-text follows the default mapping from TextBugger [50]. The number of characters for replacement is decided by a fixed ratio of the input character length, and we set the ratio to 0.05.

2) Differentiable Objective Approximation: Unlike attacking input speaker embedding, TTSlow-text encounters a non-differentiable issue since the input text is not differentiable in our optimization objective, as shown in Equation 1. Therefore, we propose to design a differentiable objective to approximate our adversarial goals.

We consider replacing the original character with another character $\hat{t}$ to achieve the optimal perturbation δ:

$$\delta = \underset{\hat{t}}{argmax}\ \text{Inc}_{t,\hat{t}}, \tag{4}$$

To compute the target character, we define character replace increment $Inc_{t,\hat{t}}$ to measure the efficiency degradation caused by replacing character t to $\hat{t}$:

$$\text{Inc}_{t,\hat{t}} = \sum_{j}(E(\hat{t}) - E(t))_j \times \frac{\partial \mathcal{L}_{\text{TTSlow-text}}(t)}{\partial t_i^j}, \tag{5}$$

where t denotes the input text token and $E(\cdot)$ represents the text embedding vector of a given token. We note that $E(\cdot)$ is differentiable, and t_i denote the i-th text token of the input text token sequence and j is the j-th dimension of the text embedding. $\text{Inc}_{t,\hat{t}}$ denotes the increase in the gradient of our objective function, resulting from replacing token t with token $\hat{t}$. $\mathcal{L}_{\text{TTSlow-text}}$ is to encourage TTS model to generate non-stop speech frames, thereby slowing down the speech synthesis process. Further details regarding the TTSlow-text loss will be provided in Section IV.

3) Perturbation Generation and Candidates Selection: Subsequently, we employ the approximated objective function to introduce minor perturbations to the input text, generating a set of adversarial candidates for both char and homo strategies that adhere to the specified imperceptibility constraints. We generate 100 adversarial candidates for both char and homo strategies.

Once the adversarial candidates are generated, we select the valid ones for the next iteration update, as illustrated in Fig. 1 (c). To accomplish this, we discard candidates that fail to meet the constraints specified in Equation 1 and then select the top three candidates based on their fitness scores for the next search iteration.

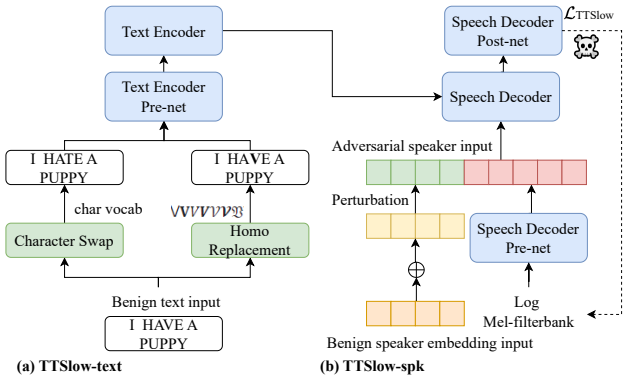


Fig. 2: Overview of TTSlow with TTSlow-text and TTSlow-spk Attacks on the autoregressive SpeechT5-tts Model

IV. TTSlow Objective Functions

Among TTS models, there are two prevalent mechanisms to determine the output speech length: length predictor for non-autoregressive models, and special end-of-speech token for autoregressive model. To apply TTSlow on a wide range of TTS models, we propose distinct objective functions for these two length control mechanisms. In this section, we introduce our proposed objective functions in detail. Specifically, we delve into an autoregressive TTS model SpeechT5 [58], and a widely adopted non-autoregressive TTS model VITS [15] to explain the attack methodologies.

A. Autoregressive TTS Model

To gain a comprehensive understanding of how TTSlow attacks operate on an autoregressive TTS model, we illustrate the attack process targeting SpeechT5-tts in Fig. 2 where Speecht5-tts processes both text and speaker embedding inputs. Its architecture comprises a text encoder pre-net, text encoder, speech decoder pre-net, speech decoder, and speech decoder post-net. The speech decoder pre-net processes the log Mel-filterbank input in an autoregressive manner [58]. Subsequently, we delve into the specifics of how TTSlow is achieved concerning speaker embedding and text input.

1) TTSlow-spk on Autoregressive TTS Model: When the given benign input is a speaker embedding, TTSlow-spk attacks the speaker embedding while keeping the text input unchanged. As depicted in Fig. 2 (b), the text encoder pre-net initially transforms the unchanged text input into an embedding vector, which is then further converted to text representation via the text encoder [58]. The X-vector [69] serves as the benign speaker embedding and is concatenated with the output of the speech-decoder pre-net, followed by a linear layer. Similarly, in Section III-C1, the adversarial speaker embedding is obtained by adding the benign speaker embedding and the perturbation, which is updated via gradient optimization.

The speech-decoder post-net comprises two modules. The first module aims to convert the decoder output to a scalar through a binary classifier for predicting the stop token [58] (scalar 0 for no-stop and 1 for stop), where TTSlow-spk is intended to encourage the prediction of non-stop tokens. The

second module predicts the log Mel-filterbank by feeding the decoder output into a linear layer and five 1-D convolutional layers [58].

To slow down the speech generation, we propose to make the victim TTS model generate longer speech by optimizing the decoder post-net output towards the conversion to no-stop tokens. Therefore, the TTSlow-spk objective in Equation 2 can be calculated as:

$$\mathcal{L}_{\text{TTSlow-spk}} = \mathcal{L}_{\text{BCE}}(f(s+\delta), y), \quad s.t. ||\delta|| \leq \epsilon, \tag{6}$$

where s denotes input speaker embedding and $\mathcal{L}_{\text{BCE}}$ represents binary cross entropy loss between the generated scalar from the binary classifier and target scalar y for non-stopping purposes.

2) TTSlow-text on Autoregressive TTS Model: When the given benign input is text, TTSlow-text attacks the text while keeping the speaker embedding input unchanged, as shown in Fig. 2 (a). The benign text input undergoes iterative modifications through differentiable objective approximation in two strategies, char, and homo, as illustrated in Section III-D2.

For instance, in Fig. 2 (a), the benign text input "I HAVE A PUPPY" is altered to "I HATE A PUPPY" by swapping the character "V" with "T" in the char strategy, and adjusted to "I HA*V*E A PUPPY" by replacing "V" with "***V***" in the homo strategy. The objective of TTSlow-text is to encourage the Victim TTS model to generate non-stop speech, and the TTSlow-text loss in Equation 5 can be computed as:

$$\mathcal{L}_{\text{TTSlow-text}} = \mathcal{L}_{\text{BCE}}(f(E(t+\delta)), y), \quad s.t. ||\delta|| \leq \epsilon, \tag{7}$$

where t denotes input text and $E(\cdot)$ represents the text embedding vector.

In summary, our proposed TTSlow approach aims to discover the optimal perturbation δ that minimizes either the loss $\mathcal{L}_{\text{TTSlow-text}}$ or $\mathcal{L}_{\text{TTSlow-spk}}$ by encouraging the generation of stop token, thereby reducing inference efficiency.

B. Non-autoregressive TTS Model

Different from autoregressive model that generates speech in a sequential manner, non-autoregressive models are able to generate speech in parallel via duration predictor. We present the TTSlow attack process on the recent state-of-the-art non-autoregressive VITS model [14], [15] in Fig. 3. VITS is a parallel end-to-end architecture based on conditional variational auto-encoder (VAE), a stochastic duration predictor for alignment generation, Transformer-based encoder, and HiFi-GAN based decoder [15].

1) TTSlow-spk on Non-autoregressive TTS Model: When provided with a benign speaker embedding, TTSlow-spk attacks the speaker embedding of VITS while keeping the text input unchanged. The adversarial speaker embedding is also obtained by adding the benign speaker embedding and the perturbation, which is updated via gradient optimization as in Section III-C1. Subsequently, the updated adversarial speaker embedding is inputted into the decoder and stochastic duration predictor. The stochastic duration predictor plays a crucial role in estimating the distribution of text token durations

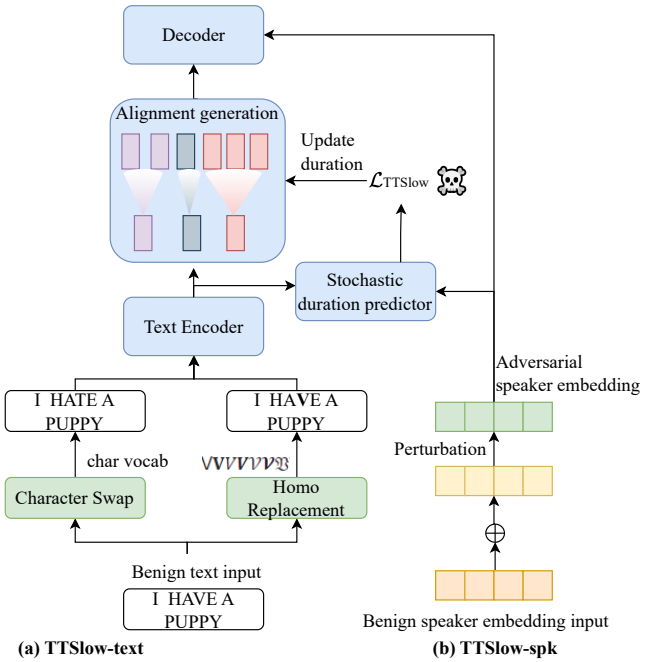


Fig. 3: Network Architecture of TTSlow with TTSlow-text and TTSlow-spk Attacks on the non-autoregressive VITS Model

and generating alignments for speech synthesis based on the inferred durations, as in Fig. 3 (b).

To this regard, we introduce a novel TTSlow-spk objective designed for non-autoregressive TTS models. The TTSlow-spk loss is formulated to maximize the cumulative output duration predicted by the stochastic duration predictor, thereby extending the length of the synthesized speech waveform. The TTSlow-spk objective, as defined in Equation 2, can be computed as follows:

$$\mathcal{L}_{\text{TTSlow-spk}} = -\sum D(s+\delta), \quad s.t. ||\delta|| \leq \epsilon, \tag{8}$$

where s denotes input speaker embedding and D signifies the output duration predicted by the stochastic duration predictor. Consequently, minimizing the TTSlow-spk loss is aimed at maximizing the output duration, ensuring continuous speech synthesis with non-stop purpose.

2) TTSlow-text on Non-autoregressive TTS Model: When provided with a benign text input, TTSlow-text attacks the text content while preserving the speaker embedding input, as depicted in Fig. 3 (a). The benign text input undergoes iterative modifications through differentiable objective approximation in two strategies, char, and homo, as illustrated in Section III-D2. We introduce an innovative objective for TTSlow-text, which incentivizes the Victim TTS model to generate longer-duration speech by amplifying the duration outputs from the stochastic duration predictor. Thus, the TTSlow-text loss defined in Equation 5 can be computed as:

$$\mathcal{L}_{\text{TTSlow-text}} = -\sum D(E(t+\delta)), \quad s.t. ||\delta|| \leq \epsilon, \tag{9}$$

where t denotes input text and $E(\cdot)$ represents the text embedding vector.

In summary, our proposed TTSlow approach aims to discover the optimal perturbation δ by optimizing either the

proposed objective $\mathcal{L}_{\text{TTSlow-text}}$ or $\mathcal{L}_{\text{TTSlow-spk}}$ via duration maximization to reduce the inference efficiency of the victim non-autoregressive TTS model.

V. Experiments

In this section, we present datasets and experimental setup.

A. Datasets

For evaluation purposes, we utilize three widely recognized TTS datasets from Huggingface: the LibriSpeech dataset [70], the LJ-Speech dataset [71], and the English dialects database [72], assessing them against multiple attack approaches. To manage computational demands, we evaluate the first 100 utterances from the LJ-Speech dataset[1] and the first 100 utterances of the clean test subset from the LibriSpeech dataset[2]. Additionally, we include the first 100 unique sentences from the Scottish female subset of the English dialects dataset [3] for further evaluation. It's important to note that the LibriSpeech and LJ-Speech datasets are English datasets, while the English dialects dataset provides English accents, contributing to the diversity of our study.

B. Experimental Setup

1) TTS Model Architecture and Attack Details: For all attack approaches, we set the total number of iterations as 100 and beam size as 3. Learning rate α is set to 0.1 for TTSlow-spk, and we refer TTSlow-spk (l2) and TTSlow-spk (linf) with ℓ_2 norm and ℓ_{inf} norm in Table I, respectively. All models are implemented in Huggingface, where SpeechT5-tts is publicly available in the link [4]. The encoder-decoder backbone in SpeechT5-tts contains twelve Transformer encoder blocks and six Transformer decoder blocks [58]. The text pre-net consists of a shared embedding layer, the speech-decoder pre-net and post-net use the same setting as in [58]. The HiFi-GAN vocoder [73] is used to convert the log Mel-filterbank to the raw waveform. Detailed parameters follow [58] and can be found in the source codes. Target scalar y is set to 0 to indicate no-stopping operation.

Our experiments utilize the publicly available VITS-VCTK model [5] as the VITS victim model for speaker-oriented attacks (referenced in Table I). This model is trained using the VITS architecture on the VCTK dataset, which consists of approximately 44,000 short audio clips spoken by 109 native English speakers, totaling around 44 hours of audio [15]. For speaker based attack approaches, we randomly select one speaker from 109 speakers to form the benign speaker embedding input. We also employ MMS-TTS model [14] as VITS victim model for text-oriented attack (also referenced in Table I), which is publicly accessible via the link [6]. MMS-TTS builds upon the VITS architecture [15] and extends its capabilities to support a triple-language setting across 1,107 languages. This advancement introduces greater challenges, given its increased power and awareness of large text corpora.

2) Baselines: We present two baseline methods for each of our proposed attack techniques. In text-oriented attacks, the text baseline involves character swaps and homo replacements, without employing the iterative differentiable approximation optimization approach designed in our method. The number of characters replaced is determined by a fixed ratio of the input character length, set at 0.05. As for the speaker embedding attack baseline (speaker baseline), we utilize Gaussian perturbations without the proposed objective function and gradient optimization techniques.

VI. Results and Discussion

To evaluate the efficacy of our attack strategies, we quantify the output speech frames (# Frames), with a higher count indicating better performance. Table I presents the maximum and mean values of this metric for each dataset, along with the mean and max increment percentages to show improvements over original counterparts. Furthermore, we conduct a comprehensive analysis of inference efficiency in Fig. 6 and Fig. 5, including factors such as inference time and inference energy consumption, using visual methods.

We assess attack performance using the attack success rate (ASR), where a higher rate indicates better performance. ASR is calculated as the ratio of successfully attacked samples (where the adversarial speech length is 20% longer than the original) to the total dataset size. Additionally, we provide adversarial samples for human imperceptibility analysis and include decoded transcriptions with their intelligibility evaluation by Gemini in Table II.

A. Vulnerability of the Victim TTS model

Our objective is to investigate the vulnerability of a TTS model to adversarial attacks. As shown in Table I, we observe that the TTS models are susceptible to all proposed adversarial attack techniques across three datasets. For instance, TTSlow-spk (l2) results in a relative 313 % increase in the length of the speech sample compared to the original clean speech sample. This finding underscores the significance of evaluating the robustness of efficiency and designing defense systems for the victim TTS model.

B. The Effectiveness of Adversarial Attacks

We assess the efficacy of the proposed adversarial attacks by comparing them to their corresponding baselines for the number of frames. As shown in Table I, both TTSlow-spk (linf) and TTSlow-spk (l2) significantly outperform the speaker baseline for all TTS models across all three datasets. Similarly, TTSlow-text consistently outperforms the text baseline for all TTS models across all three datasets. These findings demonstrate the effectiveness of the proposed attack approach.

To provide a clearer visual representation of the length distribution in our generated speech samples, we present Gaussian kernel density estimation plots for the proposed TTSlow-spk (l2), TTSlow-spk (linf), and TTSlow-text, alongside their

[1] https://huggingface.co/datasets/lj_speech
[2] https://huggingface.co/datasets/librispeech_asr
[3] https://huggingface.co/datasets/ylacombe/english_dialects
[4] https://huggingface.co/microsoft/speecht5_tts
[5] https://huggingface.co/kakao-enterprise/vits-vctk
[6] https://huggingface.co/facebook/mms-tts-eng

TABLE I: Comparison of the performance of different adversarial attack approaches on the SpeechT5 and VITS TTS models using the evaluation method, the number of frames under three datasets. Clean represents the original clean speech sample. The mean absolute value is computed by taking the average of generated speech samples for each dataset. The max absolute value is determined by selecting the highest value among # frame values of the generated speech samples for each dataset.

Evaluation	Attack Model	Datasets	Attack Methods	Mean Absolute	Max Absolute	Mean Incre	Max Incre	ASR (%)
# Frames	SpeechT5	LJ speech	Clean	105,976	159,124	0	0	0
	SpeechT5	LJ speech	Text Baseline	138,409	254,976	0.31	0.60	80
	SpeechT5	LJ speech	Speaker Baseline	90,179	180,736	-0.15	0.14	0
	SpeechT5	LJ speech	TTSlow-text	189,926	360,960	0.79	1.27	98
	SpeechT5	LJ speech	TTSlow-spk (l2)	**424,156**	**860,160**	**3.00**	**4.41**	**98**
	SpeechT5	LJ speech	TTSlow-spk (linf)	328,274	860,160	2.10	4.41	97
# Frames	SpeechT5	Librispeech	Clean	107,290	373,040	0	0	-
	SpeechT5	Librispeech	Text Baseline	141,932	527,872	0.32	0.42	64
	SpeechT5	Librispeech	Speaker Baseline	90,142	381,440	-0.16	0.02	5
	SpeechT5	Librispeech	TTSlow-text	173,562	533,504	0.62	0.43	86
	SpeechT5	Librispeech	TTSlow-spk (l2)	**443,069**	**1,817,600**	**3.13**	**3.87**	**97**
	SpeechT5	Librispeech	TTSlow-spk (linf)	394,281	1,817,600	2.67	3.87	96
# Frames	SpeechT5	English Dialects	Clean	102,222	212,992	0	0	-
	SpeechT5	English Dialects	Text Baseline	117,365	310,784	0.15	0.46	39
	SpeechT5	English Dialects	Speaker Baseline	73,344	218,112	-0.28	0.02	2
	SpeechT5	English Dialects	TTSlow-text	156,482	335,360	0.53	0.57	73
	SpeechT5	English Dialects	TTSlow-spk (l2)	**329,840**	742,400	**2.23**	2.49	**96**
	SpeechT5	English Dialects	TTSlow-spk (linf)	286,423	**773,120**	1.80	**2.63**	95
Evaluation	**Attack Model**	**Datasets**	**Attack Methods**	**Mean Absolute**	**Max Absolute**	**Mean Incre**	**Max Incre**	**ASR (%)**
# Frames	VITS	LJ speech	Clean	105,976	159,124	0	0	-
	VITS	LJ speech	Text Baseline	104,709	166,400	-0.01	0.05	8
	VITS	LJ speech	Speaker Baseline	139,616	181,674	0.32	0.14	54
	VITS	LJ speech	TTSlow-text	156,121	284,160	0.47	0.79	95
	VITS	LJ speech	TTSlow-spk (l2)	174,295	315,648	0.64	0.98	**100**
	VITS	LJ speech	TTSlow-spk (linf)	**301,839**	**744,192**	**1.85**	**3.68**	**100**
# Frames	VITS	Librispeech	Clean	107,290	373,040	0	0	-
	VITS	Librispeech	Text Baseline	105,585	320,256	-0.12	-0.14	10
	VITS	Librispeech	Speaker Baseline	99,259	132,096	-0.07	-0.65	49
	VITS	Librispeech	TTSlow-text	161,239	514,303	0.58	0.38	91
	VITS	Librispeech	TTSlow-spk (l2)	166,019	537,600	0.55	0.44	88
	VITS	Librispeech	TTSlow-spk (linf)	**264,020**	**993,024**	**1.46**	**1.66**	**98**
# Frames	VITS	English Dialects	Clean	102,222	212,992	0	0	-
	VITS	English Dialects	Text Baseline	90,071	169,216	-0.12	-0.21	9
	VITS	English Dialects	Speaker Baseline	88,399	107,562	-0.14	-0.49	26
	VITS	English Dialects	TTSlow-text	133,501	295,168	0.31	0.39	59
	VITS	English Dialects	TTSlow-spk (l2)	143,859	323,072	0.41	0.52	68
	VITS	English Dialects	TTSlow-spk (linf)	**230,044**	**544,512**	**1.25**	**1.56**	**92**

respective baselines and clean data in Fig. 4. Observing the plot, we notice that the baseline densities occupy a larger area for longer speech frames than clean, while the proposed methods' densities exhibit a significantly larger footprint compared to both the baselines and the clean data. This observation suggests that the proposed methods effectively target longer speech outputs than the baseline attack methods and the original data.

C. Attack Success Rate

To comprehensively analyze the proportion of successful attacks on a per-dataset basis, we present the attack success rate (ASR) for all the models in Table I. We observed that our proposed TTSlow achieves higher ASR values compared to their baselines. Notably, TTSlow-spk even reaches 100% ASR on LJ-Speech, indicating successful attacks on all samples. This underscores the effectiveness of our approach on TTS models across the three datasets, and identifies TTS models' weaknesses for security testing.

D. Impact of Inference Efficiency

To assess the inference efficiency of TTS models, we present Gaussian KDE plots of inference time and inference energy consumption of SpeechT5 model on LJ Speech data in Fig. 6 and Fig. 5, respectively. Note that the inference time and energy are all measured on the NVIDIA A40 GPU. We can

TABLE II: Comparison of adversarial text input samples and their decoded transcriptions from adversarially generated speech samples using Gemini, across different attacks: clean, text baseline, TTSlow-text, speaker baseline, TTSlow-spk (L2), and TTSlow-spk (Linf). Clean represents the original text of the clean speech sample.

Attack	**Adversarial text input samples**	**# Frames**
Clean	especially as regards the lower-case letters; and type very similar was used during the next fifteen or twenty years not only by Schoeffer,	154,294
Text baseline	esp*e*cially a*s* regard*s* the lower-case *l*etters; and type very similar was used dur*i*ng the next fifteen or twenty y*e*ars not only b*y* Schoeffer,	185,856
TTSlow-text	espe*c*ially as regards the lower-case letters; and type very similar w*a*s used du*r*ing the next fifteen or *t*we*n*ty years n*o*t only by S*c*hoeffer,	299,520
Speaker baseline	especially as regards the lower-case letters; and type very similar was used during the next fifteen or twenty years not only by Schoeffer,	113,152
TTSlow-spk (l2)	especially as regards the lower-case letters; and type very similar was used during the next fifteen or twenty years not only by Schoeffer,	581,632
TTSlow-spk (linf)	especially as regards the lower-case letters; and type very similar was used during the next fifteen or twenty years not only by Schoeffer,	485,888
Attack	**Decoded Transcription from Adversarially Generated Speech Samples via Gemini**	**Score**
Clean	especially as regards the lower case letters And type very similar was used during the next 15 or 20 years Not only by Schaefer	9
Text baseline	Especially regard the lower case A type very similar was used during in the next 15 or 20 years not only be show ever	6
TTSlow-text	S Ili as regards the lower case letters and type very similar double us used do in the next fifteen or twenty years in only by S Hoefer only by S Hoefer (noise) Ili as only by S Hoefer	3
Speaker baseline	especially as regards the lower case letters and type very similar was used during the next 15 or 20 years not only by Schaeffer	9
TTSlow-spk (l2)	This bell (noise) Not only holy years not only by chauffeur holy years not only chauffeur chauffeur chauffeur chauffeur	2
TTSlow-spk (linf)	especially as regards the lower case letters and type very similar was used during the next 15 or 20 years during the next 15 or 20 years not only by Schoeffer Then type very similar was used during the next 15 or 20 or 20 ors ors	5

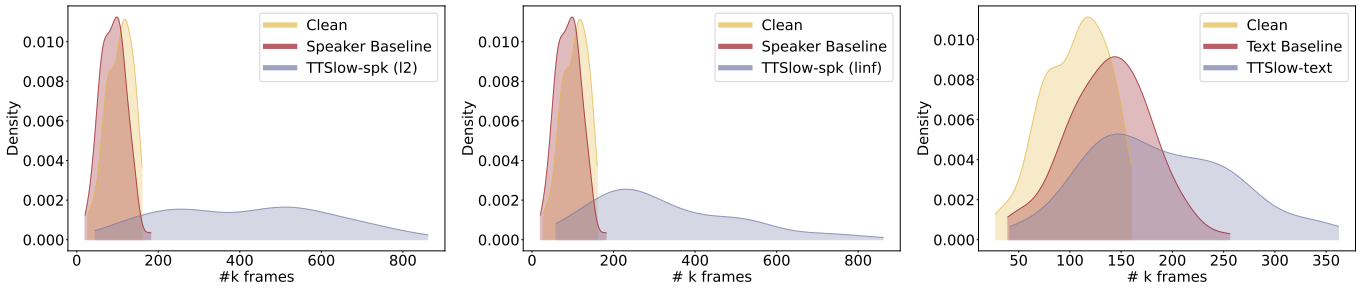


Fig. 4: Gaussian kernel density estimation (KDE) plots of speech length (# Frames) for clean data and the baseline approach with the proposed TTSlow-spk (l2), TTSlow-spk (linf), and TTSlow-text approaches on the LJ Speech dataset.

see in Fig. 5 that both the clean and baseline models consume a comparable amount of GPU energy, while the proposed TTSlow (shown in blue) significantly outperforms both the baselines and clean models. This demonstrates the effectiveness of our attack approaches in increasing inference energy and consequently reducing the overall inference efficiency of TTS models.

Similarly, in Fig. 6, we observe that the proposed TTSlow approach outperforms both the baselines and clean models significantly in terms of inference time. This suggests that TTSlow causes the TTS victim model to take longer for inference, leading to a decrease in inference efficiency. We observe from the figures that the proposed attacks not only extend the generated speech duration but also lead to higher computation, which indicates that the relevance of adversarial attacks on TTS systems lies in their potential to expose vulnerabilities and limitations within these systems.

E. Adversarial Samples

To demonstrate the impact of the proposed adversarial perturbations, we present a case study of adversarial samples generated in Table II. To enhance visibility, we utilize highlighted italic characters here to represent homoglyphs since the replacement of homoglyphs is challenging to discern with human eyes. Due to space constraints, more adversarial samples generated using TTSlow and codes can be accessed through the following link [7].

[7]https://xiaoxue1117.github.io/TTSlow/

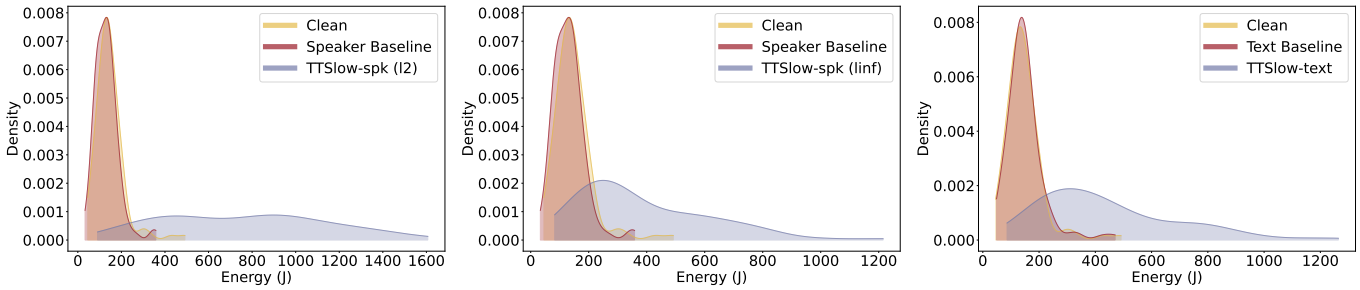


Fig. 5: Gaussian KDE plots of inference energy for clean data and the baseline approach with the proposed TTSlow-spk (l2), TTSlow-spk (linf), and TTSlow-text approaches on the LJ Speech dataset.

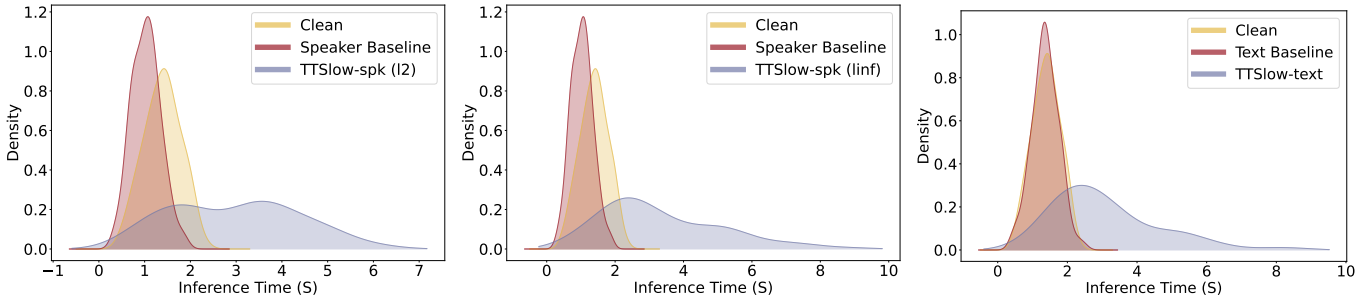


Fig. 6: Gaussian KDE plots of inference times for clean data and the baseline approach with the proposed TTSlow-spk (l2), TTSlow-spk (linf), and TTSlow-text approaches on the LJ Speech dataset.

From the provided samples, we observe seven instances of homoglyph replacements for TTSlow-text (”c”, ”a”, ”r”, ”t”, ”e”, ”o” and ”c”), resulting in approximately 61% and 94% longer speech samples compared to the text baseline and clean samples, respectively. We consider some of the attack examples are human imperceptible samples. From the samples, we observe that the proposed attacks extend the duration of generated speech while maintaining nearly human imperceptibility.

F. Evaluation with Gemini

We conducted a case study to analyze error patterns in adversarial speech samples generated by the TTSlow attacker. This analysis included an understandability evaluation and transcription analysis of the synthesized speech samples using Gemini [74], as shown in Table II. The adversarial samples were transcribed by Gemini and assigned an understandability score from 0 to 10, with 10 indicating perfect understanding and 0 indicating no understanding. The transcriptions and their scores are presented in Table II. We used Gemini 1.5 Pro, the latest large multimodal model for language, speech, and video [74].

From Table II, it is evident that the TTSlow approach yields lower understandability scores compared to clean and baseline samples. This outcome suggests that attacks successfully damage the content of the generated speech samples, making them challenging to understand. We also listened to the generated speech samples and observed error patterns, some of which can be found at the following link [8].

[8]https://xiaoxue1117.github.io/TTSlow/

Both the speech samples in the link and the decoded transcriptions in Table II revealed several error patterns. These include word repetitions (e.g., ”next 15 or 20 years” in TTSlow-spk (inf)), incorrect word generation (e.g., ”S Hoefer” in TTSlow-text), and instances of long silence or noisy speech (e.g., noise in TTSlow-spk (l2) for example 3 in the above link). This error analysis highlights that TTSlow not only reduces TTS model inference efficiency but also achieves its accuracy-oriented attack goals by rendering the generated speech content incomprehensible through these error patterns.

VII. Conclusion and Future Work

In this work, we propose TTSlow, a novel adversarial attack approach that can decrease the efficiency of both autoregressive and non-autoregressive TTS models significantly. TTSlow enables diverse attacks on both speaker embeddings and text inputs, utilizing innovative objective functions and optimization techniques that have demonstrated effectiveness. By evaluating efficiency robustness in TTS models, our work contributes to advancing TTS research, bridging the gap between no attacks and multiple attack approaches while exploring inference efficiency robustness in TTS. Through extensive experiments on three publicly available datasets, we demonstrate that the proposed TTSlow approach outperforms baselines with improved performance. This study offers valuable information for future research on the robustness of efficiency in TTS. Furthermore, this work emphasizes the critical role of audio security in TTS by showing how adversarial attacks can reveal system vulnerabilities, underscoring the need for developing more secure and resilient TTS systems for real-world applications.

In our future work, we plan to explore accuracy-oriented attacks targeting intelligibility, prosody and quality of the generated speech, along with developing advanced defense strategies to counter adversarial attacks on TTS models. Additionally, we will systematically explore the robustness of language model-based TTS systems, including recent advancements in neural codec models [75], [76] and LLM-based model [77].

VIII. Acknowledgment

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-GC-2022-005). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This research is supported by the National Research Foundation, Singapore and InfocommMedia Development Authority, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore and InfocommMedia Development Authority, Singapore.

References

[1] Yusuke Yasuda and Tomoki Toda, "Text-to-speech synthesis based on latent variable conversion using diffusion probabilistic model and variational autoencoder," in *IEEE ICASSP*, 2023, pp. 1–5.

[2] Li-Wei Chen, Shinji Watanabe, and Alexander Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," *arXiv preprint arXiv:2302.04215*, 2023.

[3] Fahima Khanam, Farha Akhter Munmun, Nadia Afrin Ritu, Aloke Kumar Saha, and Muhammad Firoz, "Text to speech synthesis: A systematic review, deep learning based architecture and future research direction," *Journal of Advances in Information Technology Vol*, vol. 13, no. 5, 2022.

[4] Rui Liu, Yifan Hu, Haolin Zuo, Zhaojie Luo, Longbiao Wang, and Guanglai Gao, "Text-to-speech for low-resource agglutinative language with morphology-aware language model pre-training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[5] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[6] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[7] Wei Ping, Kainan Peng, and Jitong Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations*, 2018.

[8] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *International Conference on Learning Representations*, 2018.

[9] Erica Cooper, Cheng-I Lai, Yusuke Yasuda, Fuming Fang, Xin Wang, Nanxin Chen, and Junichi Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6184–6188.

[10] 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*, vol. 33, pp. 8067–8077, 2020.

[11] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[12] 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," in *International Conference on Learning Representations*, 2020.

[13] Heeseung Kim, Sungwon Kim, and Sungroh Yoon, "Guided-tts: A diffusion model for text-to-speech via classifier guidance," in *International Conference on Machine Learning*. PMLR, 2022, pp. 11119–11133.

[14] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al., "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[15] 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*. PMLR, 2021, pp. 5530–5540.

[16] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[17] Yooncheol Ju, Ilhwan Kim, Hongsun Yang, Ji-Hoon Kim, Byeongyeol Kim, Soumi Maiti, and Shinji Watanabe, "Trinitts: Pitch-controllable end-to-end tts without external aligner.," in *INTERSPEECH*, 2022, pp. 16–20.

[18] Chenfeng Miao, Shuang Liang, Minchuan Chen, Jun Ma, Shaojun Wang, and Jing Xiao, "Flow-tts: A non-autoregressive network for text to speech based on flow," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7209–7213.

[19] Kainan Peng, Wei Ping, Zhao Song, and Kexin Zhao, "Non-autoregressive neural text-to-speech," in *International conference on machine learning*. PMLR, 2020, pp. 7586–7598.

[20] Aäron 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," in *9th ISCA Speech Synthesis Workshop*, pp. 125–125.

[21] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[22] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2410–2419.

[23] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[24] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[25] 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*. PMLR, 2021, pp. 8599–8608.

[26] Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren, "Prodiff: Progressive fast diffusion model for high-quality text-to-speech," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 2595–2605.

[27] Rongjie Huang, Chunlei Zhang, Yi Ren, Zhou Zhao, and Dong Yu, "Prosody-tts: Improving prosody with masked autoencoder and conditional diffusion model for expressive text-to-speech," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 8018–8034.

[28] Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[29] Tao Li, Chenxu Hu, Jian Cong, Xinfa Zhu, Jingbei Li, Qiao Tian, Yuping Wang, and Lei Xie, "Diclet-tts: Diffusion model based cross-

lingual emotion transfer for text-to-speech—a study between english and mandarin," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[30] Rui Liu, Yifan Hu, Ren Yi, Yin Xiang, and Haizhou Li, "Generative expressive conversational speech synthesis," *arXiv preprint arXiv:2407.21491*, 2024.

[31] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li, "Controllable accented text-to-speech synthesis," *arXiv preprint arXiv:2209.10804*, 2022.

[32] Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F Chen, "Emo-dpo: Controllable emotional speech synthesis through direct preference optimization," *arXiv preprint arXiv:2409.10157*, 2024.

[33] Rui Liu, Yifan Hu, Yi Ren, Xiang Yin, and Haizhou Li, "Emotion rendering for conversational speech synthesis with heterogeneous graph-based context modeling," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 18698–18706.

[34] Rui Liu, Haolin Zuo, Zheng Lian, Bjorn W Schuller, and Haizhou Li, "Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities," *IEEE Transactions on Affective Computing*, 2024.

[35] Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al., "Deep voice: Real-time neural text-to-speech," in *International conference on machine learning*. PMLR, 2017, pp. 195–204.

[36] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki, "Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, 2019.

[37] Atli Sigurgeirsson and Simon King, "Controllable speaking styles using a large language model," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10851–10855.

[38] Yinghao Aaron Li, Cong Han, Vinay S Raghavan, Gavin Mischler, and Nima Mesgarani, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *arXiv preprint arXiv:2306.07691*, 2023.

[39] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[40] Sean Vasquez and Mike Lewis, "Melnet: A generative model for audio in the frequency domain," *arXiv preprint arXiv:1906.01083*, 2019.

[41] Mirazul Haque, Rutvij Shah, Simin Chen, Berrak Şişman, Cong Liu, and Wei Yang, "Slothspeech: Denial-of-service attack against speech recognition models," *arXiv preprint arXiv:2306.00794*, 2023.

[42] Giuseppe Petracca, Yuqiong Sun, Trent Jaeger, and Ahmad Atamli, "Audroid: Preventing attacks on audio channels in mobile devices," in *Proceedings of the 31st Annual Computer Security Applications Conference*, 2015, pp. 181–190.

[43] Håkan Melin, Anna Sandell, and Magnus Ihse, "Ctt-bank: A speech controlled telephone banking system-an initial evaluation," *TMH-QPSR*, vol. 1, pp. 1–27, 2001.

[44] Kouznetsov Alexander Yu, Roman A Murtazin, Ilnur M Garipov, Anna V Kholodenina, et al., "Methods of countering speech synthesis attacks on voice biometric systems in banking," *Journal Scientific and Technical Of Information Technologies, Mechanics and Optics*, vol. 131, no. 1, pp. 109, 2021.

[45] Igor Bisio, Chiara Garibotto, Aldo Grattarola, Fabio Lavagetto, and Andrea Sciarrone, "Smart and robust speaker recognition for context-aware in-vehicle applications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8808–8821, 2018.

[46] Jiajia Liu, Shubin Zhang, Wen Sun, and Yongpeng Shi, "In-vehicle network attacks and countermeasures: Challenges and future directions," *IEEE Network*, vol. 31, no. 5, pp. 50–58, 2017.

[47] Yiming Chen, Simin Chen, Zexin Li, Wei Yang, Cong Liu, Robby Tan, and Haizhou Li, "Dynamic transformers provide a false sense of efficiency," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada, 2023, pp. 7164–7180, Association for Computational Linguistics.

[48] Yufei Li, Zexin Li, Yingfan Gao, and Cong Liu, "White-box multi-objective adversarial attack on dialogue generation," *arXiv preprint arXiv:2305.03655*, 2023.

[49] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou, "Hotflip: White-box adversarial examples for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018.

[50] J Li, S Ji, T Du, B Li, and T Wang, "Textbugger: Generating adversarial text against real-world applications," in *26th Annual Network and Distributed System Security Symposium*, 2019.

[51] Zexin Li, Bangjie Yin, Taiping Yao, Junfeng Guo, Shouhong Ding, Simin Chen, and Cong Liu, "Sibling-attack: Rethinking transferable adversarial attacks against face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24626–24637.

[52] Simin Chen, Zihe Song, Mirazul Haque, Cong Liu, and Wei Yang, "Nicgslowdown: Evaluating the efficiency robustness of neural image caption generation models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15365–15374.

[53] Raphael Olivier and Bhiksha Raj, "Recent improvements of asr models in the face of adversarial attacks," *arXiv preprint arXiv:2203.16536*, 2022.

[54] Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Almost unsupervised text to speech and automatic speech recognition," in *International conference on machine learning*. PMLR, 2019, pp. 5410–5419.

[55] Chenfeng Miao, Liang Shuang, Zhengchen Liu, Chen Minchuan, Jun Ma, Shaojun Wang, and Jing Xiao, "Efficienttts: An efficient and high-quality text-to-speech architecture," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7700–7709.

[56] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al., "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[57] Wenhao Guan, Qi Su, Haodong Zhou, Shiyu Miao, Xingjia Xie, Lin Li, and Qingyang Hong, "Reflow-tts: A rectified flow model for high-fidelity text-to-speech," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10501–10505.

[58] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al., "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 5723–5738.

[59] Simin Chen, Cong Liu, Mirazul Haque, Zihe Song, and Wei Yang, "Nmtsloth: understanding and testing efficiency degradation of neural machine translation systems," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 1148–1160.

[60] Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," *arXiv preprint arXiv:1808.05665*, 2018.

[61] Shen Wang, Zhaoyang Zhang, Guopu Zhu, Xinpeng Zhang, Yicong Zhou, and Jiwu Huang, "Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 351–364, 2022.

[62] Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, Kang Zhang, and In So Kweon, "Investigating top-k white-box and transferable black-box attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15085–15094.

[63] Yunjie Ge, Lingchen Zhao, Qian Wang, Yiheng Duan, and Minxin Du, "Advddos: Zero-query adversarial attacks against commercial speech recognition systems," *IEEE Transactions on Information Forensics and Security*, 2023.

[64] Chu-Xiao Zuo, Zhi-Jun Jia, and Wu-Jun Li, "Advtts: Adversarial text-to-speech synthesis attack on speaker identification systems," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4840–4844.

[65] Jiahe Lan, Rui Zhang, Zheng Yan, Jie Wang, Yu Chen, and Ronghui Hou, "Adversarial attacks and defenses in speaker recognition systems: A survey," *Journal of Systems Architecture*, vol. 127, pp. 102526, 2022.

[66] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *ICLR*, 2018.

[67] Raphael Olivier and Bhiksha Raj, "There is more than one kind of robustness: Fooling whisper with adversarial examples," *arXiv preprint arXiv:2210.17316*, 2022.

[68] Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot, "Bad characters: Imperceptible nlp attacks," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1987–2004.

[69] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE ICASSP*, 2018, pp. 5329–5333.

[70] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *IEEE ICASSP*, 2015, pp. 5206–5210.

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

[72] Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera, "Open-source multi-speaker corpora of the english accents in the british isles," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6532–6541.

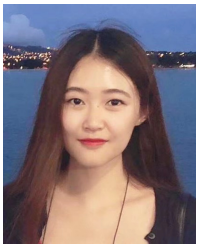
[73] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[74] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.

[75] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al., "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

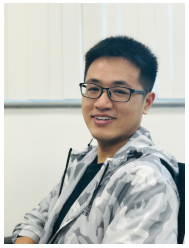
[76] Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei, "Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers," *arXiv preprint arXiv:2406.05370*, 2024.

[77] Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al., "Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens," *arXiv preprint arXiv:2407.05407*, 2024.



Xiaoxue Gao (Member, IEEE) received the Ph.D. degree from the National University of Singapore (NUS), Singapore, in 2022, and B.Eng. degree from Nanjing University, China, in 2017. She is currently a AI research scientist with the Institute for Infocomm Research (I2R), the Agency for Science, Technology, and Research (A*STAR), Singapore. She has previously worked as a postdoctoral research fellow at NUS from 2022 to 2023. Her research interests include speech recognition, self-supervised learning, large audio language models, speech synthesis, automatic lyrics transcription, audio security, and multi-modal processing.

Yiming Chen received his E.Eng. degree in Computer Science and Technology from Southern University of Science and Technology, China, in 2020. He is currently a Ph.D. student of National University of Singapore, Singapore, since 2020. His research interests include natural language processing, model safety, and efficient machine learning.



Xianghu Yue received the Ph.D. degree from the National University of Singapore (NUS), Singapore, in 2024, and B.Eng. degree in Automation from Beijing Institute of Technology, China, in 2016. He is currently a research fellow at NUS. His research interests include speech recognition, self-supervised learning, and multi-modal processing.

Yu Tsao (Senior Member, IEEE) received his B.S. and M.S. degrees in Electrical Engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and his Ph.D. degree in Electrical and Computer Engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher at the National Institute of Information and Communications Technology, Tokyo, Japan, where he worked on research and product development for automatic speech recognition in multilingual speech-to-speech translation. He is currently a Research Fellow (Professor) and the Deputy Director at the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. Additionally, he serves as a Jointly Appointed Professor in the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan, Taiwan. Dr. Tsao's research interests include assistive oral communication technologies, audio coding, and bio-signal processing. He is an Associate Editor for both IEEE Transactions on Consumer Electronics and IEEE Signal Processing Letters. In recognition of his contributions, he received the Outstanding Research Award from the National Science and Technology Council (NSTC)—the most prestigious research honor in Taiwan—in 2023. His other accolades include the Academia Sinica Career Development Award in 2017, multiple National Innovation Awards (2018–2021 and 2023), the Future Tech Breakthrough Award in 2019, the Outstanding Elite Award from the Chung Hwa Rotary Educational Foundation (2019–2020), and the NSTC FutureTech Award in 2022. Additionally, he is the corresponding author of a paper that won the 2021 IEEE Signal Processing Society (SPS) Young Author Best Paper Award.

Nancy F. Chen (Senior Member, IEEE) heads the multimodal generative AI group at A*STAR. She has served as the program chair of ICLR 2023, 2023 IEEE SPS Distinguished Lecturer, ISCA Board Member (2021-2025), in addition to being listed as 100 Women in Tech in Singapore 2021. Her honors include A*STAR Fellow (2023), the 2020 Procter & Gamble (P&G) Connect + Develop Open Innovation Award, the 2019 L'Oréal Singapore For Women in Science National Fellowship, Outstanding Mentor Award from the Ministry of Education in Singapore (2012), the Microsoft-sponsored IEEE Spoken Language Processing Grant (2011), and the NIH (National Institute of Health) Ruth L. Kirschstein National Research Award (2004-2008) and best paper awards from EMNLP, MICCAI, SIGDIAL, APSIPA, IEEE ICASSP. Speech evaluation technology developed by her team is deployed at the Ministry of Education in Singapore to support home-based learning during the COVID-19 pandemic. Nomopai is a spin-off company that uses technology from her lab to make customer agents more confident and empathetic. Prior to working at A*STAR, Dr. Chen worked at MIT Lincoln Laboratory on multilingual speech processing and obtained her PhD from MIT and Harvard University.