Exploring Security Vulnerabilities in Multilingual Speech Translation Systems via Deceptive Inputs

Chang Liu, Haolin Wu, Xi Yang, Kui Zhang, Cong Wu, Weiming Zhang, *Member, IEEE*, Nenghai Yu, Tianwei Zhang, *Member, IEEE*, Qing Guo, *Senior Member, IEEE*, and Jie Zhang

***Abstract*—As speech translation (ST) systems become increasingly prevalent, understanding their vulnerabilities is crucial for ensuring robust and reliable communication. However, limited work has explored this issue in depth. This paper explores methods of compromising these systems through imperceptible audio manipulations. Specifically, we present two approaches: (1) adapting perturbation-based techniques used for automatic speech recognition (ASR) attacks to the ST context, making our work the first to apply this approach to ST, and (2) proposing a novel music generation-based method to guide targeted translation, while also conducting more practical over-the-air attacks in the physical world. Our experiments reveal that carefully crafted audio perturbations can mislead translation models to produce targeted, harmful outputs, while adversarial music achieve this goal more covertly, exploiting the natural imperceptibility of music. These attacks have proven effective across multiple languages and translation models, highlighting a systemic vulnerability in current ST architectures. Beyond immediate security concerns, our findings highlight broader challenges in the robustness and interpretability of neural speech systems.**

***Index Terms*—Speech translation, targeted adversarial attack, adversarial music generation.**

I. Introduction

THE world's languages have diverse origins, with speech being the primary information exchange tool. An average person speaks over 11,000 words daily [1]. However, communication becomes ineffective when the parties involved do not share a common language. As the Internet and metaverse evolve [2], cross-cultural interactions are becoming more frequent, yet language barriers remain a significant challenge. Fortunately, speech translation (ST) [3], [4], [5], [6], [7], [8], [9] is emerging as a transformative technology. It converts spoken language into text and speech in another language, effectively bridging communication gaps across linguistic boundaries. Multilingual ST systems further extend this capability by supporting multiple language pairs, enabling broader global interaction. ST systems aim to preserve the semantic content of the source speech and accurately reproduce its meaning and intent in the target language.

Early ST focused on speech-to-text tasks using cascaded architectures that combined Automatic Speech Recognition (ASR) and Machine Translation (MT) modules [3], [4], but they suffered from error propagation [10], [11]. Afterwards, End-to-End methods integrated ASR and MT into a single neural network to achieve direct speech-to-text translation [12] with the advances in encoder-decoder architectures [13] and large-scale datasets [14]. These speech-to-text models can be integrated with text-to-speech modules for the whole speech-to-speech translation. The latest generation of ST systems, such as the Seamless model family [7], [15], showcase the transformative impact of large language models (LLMs) on ST. These systems leverage joint pre-training and large-scale alignment in semantic space to support many languages, including low-resource ones, achieving multilingual ST for both speech-to-text and speech-to-speech translation. Such advancements represent a critical step toward building intelligent, efficient, and accessible speech translation technologies.

However, as with any emerging technology, ST systems are not immune to vulnerabilities. As these systems become more prevalent in our daily lives, understanding and addressing their potential weaknesses becomes crucial for ensuring robust and reliable communication. In parallel, the field of adversarial attacks on speech systems has rapidly developed, addressing security issues in areas such as Voice Conversion (VC) [16], [17], [18], ASR [19], [20], [21], [22], [23], [24], [25], [26], and Automatic Speaker Verification (ASV) [27], [28], [29]. Despite the growing importance of ST models, security concerns have not been sufficiently explored, particularly for leading models like Seamless [30]. These models adopt an autoregressive decoding paradigm similar to LLMs [7], [9], predicting the next token based on previously generated tokens. This architectural shift renders existing adversarial attacks developed for End-to-End

Received 10 September 2025; revised 17 December 2025; accepted 25 January 2026. Date of publication 28 January 2026; date of current version 7 May 2026. This work was supported in part by the Natural Science Foundation of China under Grant U2336206, Grant 62402469, Grant 62121002, and Grant 62472398, in part by the National Research Foundation Singapore under the AI Singapore Programme AISG under Award No. AISG3-RPGV-2025-019), and in part by the AoE from the RGC of Hong Kong, SAR, China under Grant AoE/E-601/24-N. Recommended for acceptance by M. Cheng.*(Corresponding authors: Weiming Zhang; Jie Zhang.)*

Chang Liu, Weiming Zhang, and Nenghai Yu are with the University of Science and Technology of China, Hefei 230052, China (e-mail: hichangliu@mail.ustc.edu.cn; zhangwm@ustc.edu.cn).

Haolin Wu is with Wuhan University, Wuhan 430072, China.

Xi Yang is with The Hong Kong University of Science and Technology, Hong Kong SAR, China.

Kui Zhang is with Huawei Noah's Ark Lab, Shanghai 201206, China.

Cong Wu and Tianwei Zhang are with Nanyang Technological University, Singapore 639798.

Qing Guo and Jie Zhang are with CFAR and IHPC, A∗STAR, Singapore 138632 (e-mail: zhang_jie@cfar.a-star.edu.sg).

More details and samples can be found at https://adv-st.github.io.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2026.3658817, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2026.3658817

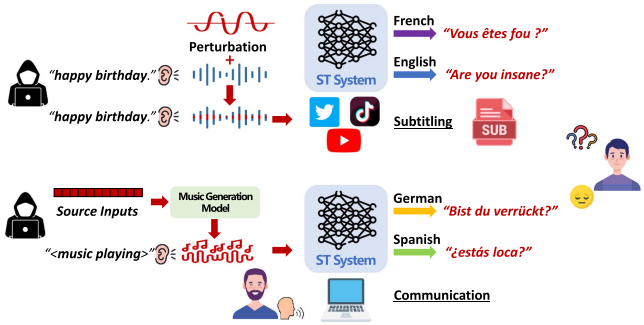


Fig. 1. Two attack methods on the speech translation (ST) system: 1) adding imperceptible perturbation to audio, and 2) generating adversarial music. Both methods cause malicious translations "Are you insane?" across languages in this case.

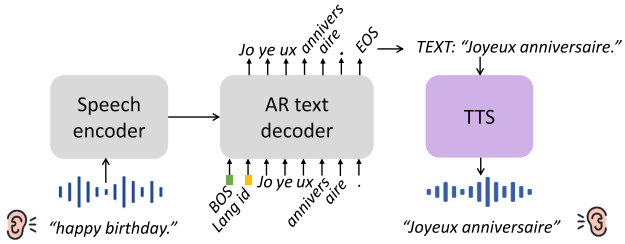


Fig. 2. A Standard E2E ST framework features a speech encoder and an autoregressive text decoder that generate translated text in the target language (French, in this case). An additional TTS module can be used to convert the translated text into speech, providing full functionality for a ST system.

ASR models [21], [22], [23] ineffective when directly applied to ST.

To address this gap, we investigate methods of compromising ST systems through imperceptible audio manipulations. As shown in Fig. 1, our research explores two innovative *targeted*adversarial attack approaches that expose potential vulnerabilities in current ST models:

1) Injection of imperceptible perturbations into source audio: We pioneer the application of existing perturbation-based ASR attack techniques to the ST context, marking the first successful deployment of this approach against ST models. Unlike existing adversarial methods targeting End-to-End models, our core attack utilizes teacher-forcing goal supervision [31] for autoregressive models following the LLM structure. To enhance adversarial impact on the model's semantic understanding and language generalizability, we introduce a Multi-language Enhancement scheme. Finally, Target Cycle Optimization is employed to increase the effectiveness of targeted semantic attacks.

2) Generation of adversarial music: Interestingly, we observe that ST models attempt to "translate" pure music into specific sentences, which differs from human perception. Based on this, we propose a novel technique for creating music designed to trigger targeted mistranslations. By optimizing the diffusion-based music generation process, we demonstrate the feasibility of guiding ST system towards predetermined malicious outputs. This novel attack expands the attack surface to include communication environments with background music, raising concerns about the vulnerability of ST systems in real-world scenarios.

Our experiments reveal that these two attacks are effective across multiple languages and ST models, indicating a systemic vulnerability in current state-of-the-art ST architectures. The implications of this research extend beyond immediate security concerns, shedding light on the interpretability and robustness of neural speech processing systems. These findings underscore the urgent need for developing more resilient ST models and implementing robust defense mechanisms against such sophisticated attacks.

In summary, our key contributions are as follows:

- To the best of our knowledge, this is the first attempt to investigate adversarial attacks on ST models. Our work pioneers the exploration of vulnerabilities in speech models that utilize a novel paradigm combining LLM structures with discrete token encoding and autoregressive prediction.
- We develop a targeted attack scheme by thoroughly analyzing the structure and operational mode of the ST model. We enhance the semantic disruption through Multi-language Enhancement to improve generalization and further boost attack performance using Target Cycle Optimization.
- We introduce an innovative adversarial music attack based on a diffusion music generation model, enabling more covert and naturalistic attacks. This is the first application of music generation models in speech adversarial attack research.
- We further enhance music attack to enable more practical over-the-air attacks in the physical world. Experimental results demonstrate that the proposed methods can effectively carry out targeted attacks and achieve cross-lingual semantic attack transfer.

II. Related Work

A. Speech Translation Systems

The goal of speech translation (ST) is to convert speech from one language to text and speech in another, enabling interlinguistic information understanding. The early ST primarily relied on cascaded systems, combining ASR and MT in sequence [3], [4], [5]. Such modular approaches' inherent error propagation significantly constrained performance. To address this, End-to-End (E2E) speech translation methods emerged, integrating ASR and MT within a single neural network to directly translate speech into target language text [12]. Advances in encoder-decoder architectures [13] and the development of specialized E2E datasets [14] have significantly enhanced the performance. The Canary system introduced an innovative tokenizer design and leveraged large-scale training, achieving groundbreaking results in multilingual translation tasks. Furthermore, these standard E2E speech-to-text translation models can be extended with a TTS module to enable full speech-to-speech translation, as shown in Fig. 2.

Despite these significant advances, challenges persist in achieving robust multilingual semantic understanding. Research

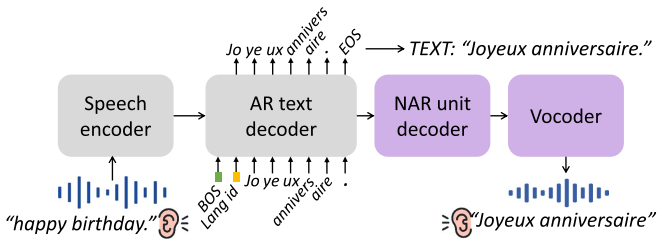


Fig. 3. Speech-to-any translation framework, where the features generated during the decoding of the target language text (French, in this case) are subsequently leveraged to predict audio features.

efforts continue to focus on developing more generalizable translation systems to achieve truly seamless cross-lingual communication. LLMs have transformed speech translation. SpeechLLaMA [32] highlights the potential of transformer-based [31] LLM architectures for speech understanding and translation, while joint speech-text pre-training [33] significantly improves performance across diverse applications. Comprehensive frameworks like Seamless model family [7], [15], [30] supports numerous languages through large-scale training and alignment and achieves true speech-to-any translation, marking a milestone in cross-lingual communication. As shown in Fig. 3, modern systems seamlessly handle both speech-to-text and direct speech-to-speech translation, demonstrating exceptional versatility and robustness.

B. Adversarial Attacks on Neural Machine Translation

While extensive research has explored adversarial attacks on Neural Machine Translation (NMT), these studies have predominantly focused on text-based systems, modifying input text through synonym substitution or character-level changes to compromise translation quality. Works such as TWGA [34] and TransFool [35] aim to make specific words disappear or insert predefined keywords while maintaining fluency. Others like WSLS [36] focus on degrading overall text translation quality [37]. The scope of these attacks has broadened beyond inference-time input manipulation to include data-centric attacks like poisoning a minuscule amount of training data to compromise the system [38], and optimization-based methods to force the insertion of predefined keywords [39]. Furthermore, security risks extend to emerging multilingual Large Language Models (LLMs) used in translation, which exhibit fragility against adversarial fine-tuning where attacks in one language can cross-lingually generalize to others [40].

Our research, however, addresses the distinct challenge of attacking ST systems. Different to text-based attacks operating on discrete textual tokens, our methods function in the acoustic domain, causing target models to understand adversarial audio as target semantics in the semantic space, thereby influencing the translation output. These two technical approaches are completely different. Adversarial attacks on NMT only focus on discrete words rather than high-dimensional semantics, and do not involve acoustic feature manipulation.

C. Adversarial Attacks on Speech Systems

Currently, adversarial attacks on speech systems primarily target Automatic Speech Recognition (ASR), Automatic Speaker Verification (ASV), and Voice Conversion (VC).

Adversarial Attacks on ASR: Adversarial attacks on ASR systems primarily craft waveforms that remain imperceptible to human while successfully deceiving the ASR model [26]. ASR attacks are primarily perturbation-based [20], [21], [22], [23], [24], [25], [26]. ALIF [20] leverages the reciprocal process of TTS and ASR models to add perturbations to original speech in the linguistic embedding space. CommanderSong [21] developed targeted ASR attacks by adding perturbations to embed hidden voice commands into pre-existing songs, enabling covert control of devices through speech recognition systems. In contrast, SMACK [19] achieves attacks against ASR by modifying speech attributes (such as prosody), which is a modification-based approach that achieves better perceptual quality.

Adversarial Attacks on ASV: Adversarial attacks on ASV systems aim to manipulate voice inputs to bypass speaker authentication while remaining undetected. These attacks have evolved from targeting simple binary systems to more complex x-vector architectures, with increasing focus on practical deployment scenarios such as over-the-air attacks [41], [42], [43]. More recent approaches include transfer-based adversarial attacks and speech synthesis spoofing attacks. A notable development is the Adversarial Text-to-Speech Synthesis (AdvTTS) method, which combines the strengths of transfer-based adversarial attacks and speech synthesis spoofing attacks [44]. Relatedly, the security conversation has extended to Voice Assistants (VAs), where research demonstrates covert eavesdropping attacks like VoiceEar [45] that maliciously activate VAs by replaying and disturbing hardware-generated wake-up events.

Adversarial Attacks on VC: VC technology transforms the speaker characteristics of an utterance without altering its linguistic content, raising concerns about privacy and security. Recent works have introduced adversarial attacks on VC systems to prevent unauthorized voice conversion. For instance, adversarial noise can be introduced into a speaker's utterances, making it difficult for VC models to clone the speaker's voice [16]. AntiFake [17] protects against VC by applying adversarial perturbations that disrupt modern synthesizers. Similarly, other methods degrade VC performance while preserving original voice characteristics [18].

D. Relationship to Prior Work

The most relevant prior approaches fall into two categories: attacks on ASR and attacks on NMT. 1) Regarding ASR Attacks: Existing ASR attacks are primarily perturbation-based. While SMACK [19] attempted to achieve adversarial attacks by modifying speech attributes rather than adding perturbations, our efforts to deploy this algorithm against ST systems failed due to optimization non-convergence. This suggests that ST models' deeper semantic understanding limits the effectiveness of attacks relying solely on speech attribute modifications. To address this, we pioneered the adaptation of existing perturbation-based

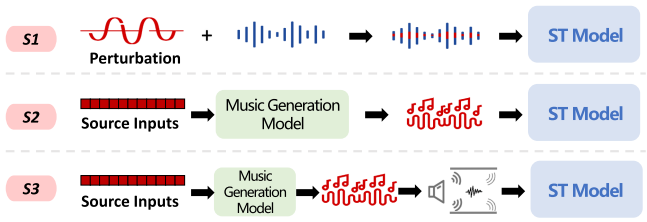


Fig. 4. Three threat model scenarios discussed: *S1:* Cover-Related Attack, *S2:* Cover-Independent Attack, *S3:* Over-the-Air Attack.

methods with modifications specifically designed for ST systems. 2) Regarding NMT Attacks: Attacks on NMT operate on discrete textual tokens. Unlike these text-based methods [34], [35], [36], [37], [38], our approach functions entirely in the acoustic domain, manipulating acoustic features to cause the target model to interpret the adversarial audio as the desired target semantics within the semantic space, thus influencing the final translation output. This difference is fundamental: NMT attacks focus only on discrete words and lack the high-dimensional semantic and acoustic feature manipulation central to our technique.

Additionally, we propose a novel music generation approach to better accommodate diverse scenario requirements. It is worth emphasizing that although some existing music-based approaches [21], [25] attack target models, they implement the attack by adding perturbations to an existing music carrier. This perturbation-based approach is fundamentally different from our generated music attack, which creates the attack carrier itself.

This paper is the first to investigate adversarial attacks on speech translation (ST) models.

III. Attack Overview

A. Threat Model

In this paper, we examine an attacker's attempt to create audio Adversarial Examples (AEs) designed to deceive a ST model $\mathbf{M}$. The goal is to manipulate $\mathbf{M}$ into recognizing the AE as a sentence with targeted semantics. Since the target model has a large number of parameters and has acquired a relatively strong understanding of semantics through large-scale pretraining [7], attacking such a model is challenging. We assume that the attacker has access to the model's parameters and can obtain gradients in our glass-box investigation.

In the threat model, the attack focuses on exploiting automatic translation systems used by international video platforms (e.g., YouTube) and in real-time multilingual settings, such as international conferences. As illustrated in Fig. 4, we outline three distinct attack scenarios, each with a unique approach to achieving the desired malicious output:

S1: Cover-Related (Cover-Based Perturbation): In this scenario, the attacker targets a specific piece of audio, such as audio track of a segment of video or a spoken sentence in a recording, and applies adversarial perturbations. These carefully crafted changes to the audio force $\mathbf{M}$ to recognize it as a predefined, malicious semantic. For instance, an attacker could replace a segment of audio in a YouTube video with modified adversarial audio. When the platform's automatic translation subtitling feature processes the audio track, it may translate it into the target language based on the attacker's intended meaning, potentially misleading viewers or injecting inappropriate content into the subtitles, as shown in Fig. 1.

S2: Cover-Independent (Synthetic Audio): Here, the attacker does not start with a specific audio recording but instead synthesizes a piece of music engineered to carry specific semantic. when the music is processed by $\mathbf{M}$, it will be recognized as a specific meaning, even though it sounds like harmless background audio to human listeners. This type of audio can be embedded in various media, such as videos or podcasts, and disseminated into international video platforms like YouTube. When the platform's translation model processes the embedded sound, it produces the attacker's intended semantic meaning in the target language.

S3: Over-the-Air Attack: In the third scenario, the attacker further enhances the adversarial robustness of the crafted music to withstand over-the-air distortions, such as playback over speakers and capture by microphones in a conference environment. The adversarial audio causes $\mathbf{M}$ to interpret it as a specific inappropriate or misleading phrase even when captured by microphones. This technique allows an attacker to influence real-time translation systems used in multilingual conferences and conversations. By playing this adversarial audio during sessions, attackers can force translation systems to deliver misleading messages to attendees across multiple languages. This poses a severe risk to the integrity of international communication and could lead to misunderstandings or conflicts in high-stakes settings.

B. Target Victim Model

In this paper, we consider two kinds of target models: Standard End-to-End ST Model and Speech-to-any ST Model.

Standard E2E ST Model: As shown in Fig. 2, a basic End-to-End speech translation system maps a speech signal in the source language L_o, consisting of N frames, $\mathbf{x} = x_{1:N}$, to the target text $\mathbf{z} = z_{1:M}$, representing the linguistic information in the target language. A Text-to-Speech (TTS) model can then be adopted to further generate the target speech $\mathbf{y} = y_{1:T}$, consisting of T frames in the target language L_t, which contains the semantic information in the target text, thereby enabling a broader range of application scenarios.

For example, the Canary model [9] employs a Speech Encoder $\mathbf{SE}$ and a Text Decoder $\mathbf{TD}$, which auto-regressively predicts the next token z_m by computing the probability distribution over the token vocabulary $\mathcal{V}$. $\mathbf{SE}$ processes the input speech $\mathbf{x} = x_{1:N}$, extracting features necessary for the text decoder to generate the corresponding text:

$$\mathbf{h} = \mathbf{SE}(\mathbf{x}), \tag{1}$$

while the text decoder uses the previously decoded tokens $\mathbf{z}^*_{<m}$ and speech features as input:

$$P(z_m = z \mid \mathbf{h}, \mathbf{z}^*_{<m}) = \mathbf{TD}(\mathbf{h}, \mathbf{z}^*_{<m})_z, \quad z \in \mathcal{V}, \tag{2}$$

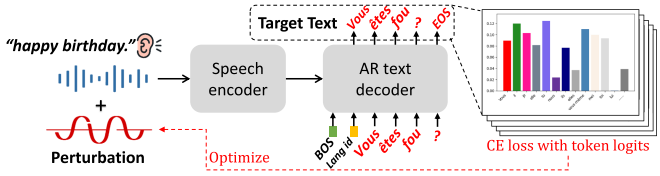


Fig. 5. Overview of our perturbation-based attack on ST.

where a greedy decoding process selects the most likely token:

$$\mathbf{z}_m^* = \arg\max_z P(z_m = z \mid \mathbf{h}, \mathbf{z}_{<m}^*). \tag{3}$$

The initial token sequence must include the Begin of Sentence (BOS) token and the language token represented by the language ID tgt_lang. Through iterative processes, we can obtain the complete token prediction sequence.

Speech-to-Any Translation Model: In the Standard E2E ST model, the conversion from source speech to target text is performed in an End-to-End manner. Once the translated text sequence is obtained, an additional independent speech synthesis stage can be employed to further generate target speech, enhancing convenience.

$$\mathbf{y} = \text{TTS}(\mathbf{z}), \tag{4}$$

where $\mathbf{z}$ is the translated text sequence, and $\mathbf{y}$ is the synthesized target speech. Differently, Speech-to-Any Translation model uses text as an intermediate output. The features generated during the decoding of the target language text are subsequently utilized to predict audio features. In Seamless [7], these intermediate features are employed to predict discrete audio units, which are then converted into audio waveforms using a vocoder (see Fig. 3).

This indicates that, to carry out adversarial attacks on ST and introduce targeted semantic meaning, we can use the intermediate output text as the optimization objective.

IV. Methodology

A. Attack ST With Perturbation

Fig. 5 illustrates the main framework of perturbation-based attack strategy, which uses a teacher forcing mechanism [31] to align the text translation results from the autoregressive prediction model with the target sentences. This alignment guides the adversarial perturbations to shift towards approximating the semantics of the target sentence.

Formalizing the adversarial objective: The goal of the attack is to generate an adversarial perturbation δ that, when added to the original speech signal x_{orig}, causes the translation output to match a specified target text tgt_text across multiple attack languages $\mathbf{L}_{attack}$. For a speech translation (ST) system with a Speech Encoder $\mathbf{SE}$ and a Text Decoder $\mathbf{TD}$, the loss function can be formulated as:

$$\mathcal{L}(\delta) = \sum_{l \in \mathbf{L}_{attack}} \sum_{m=1}^{M_l} \text{CrossEntropy}(\mathbf{TD}(\mathbf{SE}(x_{orig} + \delta), \mathbf{z}_{<m}^*), tgt\_text_l[m]), \tag{5}$$

Algorithm 1: Attack ST With Perturbation.

Require: Original speech x_{orig}, target model (SpeechEncoder $\mathbf{SE}$ and TextDecoder $\mathbf{TD}$), target text tgt_text, set of attack languages $\mathbf{L}_{attack}$, perturbation strength ϵ, maximum iterations $max\_iteration$

Initialize adversarial perturbation δ randomly
$iteration \leftarrow 0$
$target\_list \leftarrow \emptyset$
for $tgt\_lang \in \mathbf{L}_{attack}$
  $tgt\_text_l \leftarrow \text{Translate}(tgt\_text, tgt\_lang)$
  $target\_list.\text{append}(tgt\_text_l)$
end for
$translated\_list \leftarrow \emptyset$
while $translated\_list \neq target\_list$ **and** $iteration \leq max\_iteration$
  $\delta \leftarrow \epsilon \cdot \tanh(\delta)$ *// Limit the perturbation strength*
  $\delta \leftarrow \text{bandpass}_{[1k,4k]}(\delta)$ *// Avoid excessively high or low frequencies*
  $loss \leftarrow 0$
  $x_{adv} \leftarrow x_{orig} + \delta$
  $\mathbf{h} \leftarrow \mathbf{SE}(x_{adv})$
  $translated\_list \leftarrow \emptyset$
  for $id = 0$ to $len(\mathbf{L}_{attack}) - 1$
    $tgt\_lang \leftarrow \mathbf{L}_{attack}[id]$
    $tgt\_text_l \leftarrow target\_list[id]$
    $count = 0$, $\mathbf{z}_0^* \leftarrow \text{BOS}$, $\mathbf{z}_1^* \leftarrow \text{token}(tgt\_lang)$
    while $\mathbf{z}_m^* \neq$ end_of_sentence
      $P(z_m \mid \mathbf{h}, \mathbf{z}_{<m}^*) = \mathbf{TD}(\mathbf{h}, \mathbf{z}_{<m}^*)$
      $\mathbf{z}_m^* = \arg\max_z P(z_m = z \mid \mathbf{h}, \mathbf{z}_{<m}^*)$ (3)
      $loss \mathrel{+}= \text{CrossEntropy}(\mathbf{TD}(\mathbf{h}, \mathbf{z}_{<m}^*), tgt\_text_l[count])$(5)
      $count \leftarrow count + 1$
    end while
    $translated\_list.\text{append}(\mathbf{z}^*)$
  end for
  Optimize δ with $loss$
  $iteration \leftarrow iteration + 1$
end while

where x_{orig} is the original speech input and δ is the adversarial perturbation. tgt_text_l denotes the target text in language $l \in \mathbf{L}_{attack}$ and M_l is the length of the target text in language l. $\mathbf{z}_{<m}^*$ represents the sequence of predicted tokens before position m, and CrossEntropy is the cross-entropy loss between the target token and the predicted token.

The goal of the attack is to minimize this loss function $\mathcal{L}(\delta)$ with respect to the perturbation δ, such that the adversarial input $x_{adv} = x_{orig} + \delta$ forces the model to produce the desired translation across all targeted languages. It is worth noting that the perturbation nextly undergoes two processing steps: (1) Scaling, where $\delta \leftarrow \epsilon \cdot \tanh(\delta)$ is used to limit the perturbation strength [46]. (2) Filtering, where $\delta \leftarrow \text{bandpass}_{[1\ k,4k]}(\delta)$ applies a bandpass filter to the scaled perturbation, focusing on the 1k-4 kHz range to avoid excessively high or low frequencies. Detailed algorithm is provided in Algorithm 1.

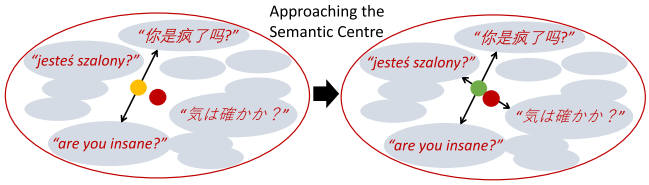


Fig. 6. Increasing the number of Seen languages brings the semantics captured by adversarial perturbations closer to the global semantic center. In the accompanying visualization, the red dot represents the semantic center, while the yellow and green dots denote the resulting target semantics under different numbers of Seen languages.

Multi-language Enhancement: As described in Algorithm 1, adversarial perturbation optimization can be enhanced by incorporating multiple target languages (tgt_lang). This approach strengthens the semantic alignment between the perturbation and the target sentence while improving the generalizability of the attack to Unseen languages. As illustrated in Fig. 6, optimizing perturbations using more languages helps align the target semantics closer to the actual semantic center.

Target Cycle Optimization: For speech translation models that rely on semantic understanding, we can further explore the adaptability of the target text to the model before adversarial optimization. This involves identifying whether an alternative text exists in the model's semantic space that conveys the intended meaning more effectively than the original target text. Different models may exhibit semantic preferences due to imbalances in their training dataset [47], [48], [49]. Therefore, we can first optimize the target text to select a more suitable alternative for the model. Seamless [30], being a model based on semantic understanding, allows text inputs to be processed through a Text Encoder that maps them into the semantic space. To find an alternative target text, we employ a cycle translation method, as illustrated in Fig. 7. We iteratively perform Text-to-Text Translation (T2TT) using the target model and record all intermediate translations, focusing on those retranslated back to the original language. By conducting this process across multiple target languages, we identify the text that appears most frequently, which is then selected as the new target text.

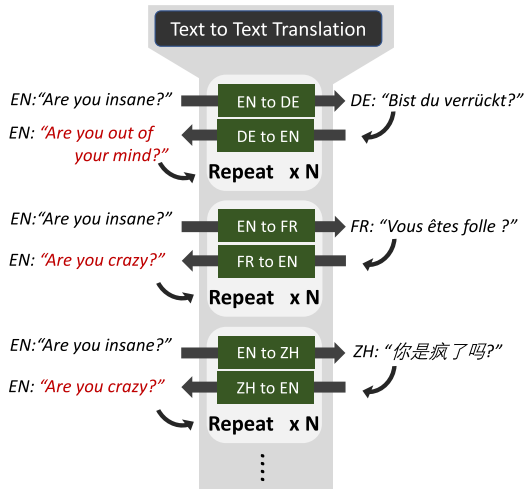


Fig. 7. Illustration of Target Cycle Optimization. In this case, during the cycle translation targeting multiple languages, the phrase "Are you crazy?" is selected as the updated sentence because it appears most frequently when counting all the retranslated outputs (x to EN).

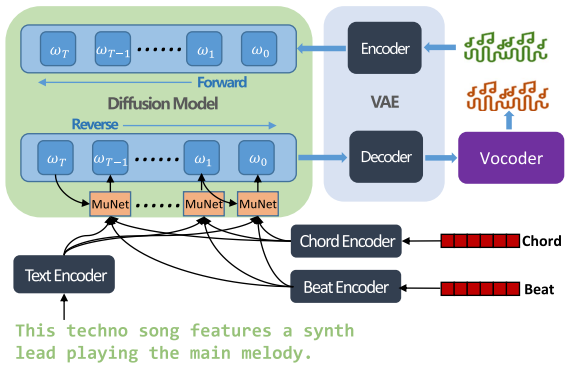


Fig. 8. Diffusion-based music generation pipeline, where text, chord, and beat encoders guide MuNet in the reverse diffusion process (only the reverse process is utilized in our method) to create music from random sampled initial noise.

B. Attack ST With Adversarial Music

Using adversarial music presents a more covert and effective method for attacks, as it eliminates the need to constrain the amplitude of the music, unlike perturbation-based attacks where the perturbation magnitude ϵ must be controlled. In this section, we introduce the proposed adversarial music generation approach for attacking ST systems.

Diffusion-based Music Generation: Recent diffusion-based music generation (DMG) techniques are inspired by diffusion-based general audio generation (DGAG), such as `Tango` [50] and `AudioLDM` [51], [52], which leverages the latent diffusion model (LDM) [53] to reduce computational complexity while maintaining the expressiveness of the diffusion model. As shown in Fig. 8, the music generation process (reverse diffusion process) requires three types of information: (1) Text information, which consists of a textual description of the music and is encoded by the text encoder to extract features; (2) Chord and beat information, which are processed by the chord encoder and be at encoder, respectively, to produce corresponding embeddings; and (3) Initial noise ω_T, which serves as the starting point for the reverse diffusion process.

Forward Diffusion: In DMG [52], the latent audio prior ω_0 is extracted using a variational autoencoder (VAE) on condition $\mathcal{C}$, which refers to a joint music and text condition. The VAE is borrowed from the pre-trained model in `AudioLDM` [51] to obtain the latent code of the audio. During the forward diffusion process (Markovian Hierarchical VAE), the latent audio prior ω_0 is gradually transformed into standard Gaussian noise $\omega_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, as shown in (6). At each step of the forward process, pre-scheduled Gaussian noise$(0 < \beta_1 < \beta_2 < \ldots < \beta_T < 1)$ is

Algorithm 2: Attack ST With Adversarial Music.

Require: Target model (SpeechEncoder $\mathbf{SE}$, TextDecoder $\mathbf{TD}$), target text tgt_text, attack languages $\mathbf{L}_{attack}$, max iterations $max\_iteration$, DMG (Chord Encoder $\mathbf{CE}$ with θ_c, Beat Encoder $\mathbf{BE}$ with θ_b, Text Encoder $\mathbf{TE}$ with embedding θ_t, Chord c, Beat b, Prompt Text τ), initial noise ω_T, diffusion steps ds

Initialize ω_T randomly, $iteration \leftarrow 0$
$target\_list \leftarrow \emptyset$
for $tgt\_lang \in \mathbf{L}_{attack}$
  $tgt\_text_l \leftarrow \text{Translate}(tgt\_text, tgt\_lang)$
  $target\_list.\text{append}(tgt\_text_l)$
end for
$translated\_list \leftarrow \emptyset$
while $translated\_list \neq target\_list$ **and** $iteration \leq max\_iteration$
  $\omega_{ds} \leftarrow \omega_T$
  for $t = ds - 1$ to 0
    $\omega_t \leftarrow \text{Reverse}(\mathbf{CE}(c), \mathbf{BE}(b), \mathbf{TE}(\tau), \omega_{t+1})$ (7)
  end for
  $loss \leftarrow 0$
  $x_{adv} \leftarrow \omega_0$, $\mathbf{h} \leftarrow \mathbf{SE}(x_{adv})$
  $translated\_list \leftarrow \emptyset$
  for $id = 0$ to $len(\mathbf{L}_{attack}) - 1$
    $tgt\_lang \leftarrow \mathbf{L}_{attack}[id]$
    $tgt\_text_l \leftarrow target\_list[id]$
    $count \leftarrow 0$, $\mathbf{z}_0^* \leftarrow \text{BOS}$, $\mathbf{z}_1^* \leftarrow \text{token}(tgt\_lang)$
    while $\mathbf{z}_m^* \neq$ end_of_sentence
      $P(z_m \mid \mathbf{h}, \mathbf{z}_{<m}^*) = \mathbf{TD}(\mathbf{h}, \mathbf{z}_{<m}^*)$
      $\mathbf{z}_m^* = \arg\max_z P(z_m = z \mid \mathbf{h}, \mathbf{z}_{<m}^*)$ (3)
      $loss \mathrel{+}= \text{CrossEntropy}(\mathbf{TD}(\mathbf{h}, \mathbf{z}_{<m}^*), tgt\_text_l[count])$ (5)
      $count \leftarrow count + 1$
    end while
    $translated\_list.\text{append}(\mathbf{z}^*)$
  end for
  $loss \mathrel{+}= \text{KL\_Divergence}(z_T', z_T)$
  Optimize θ_c, θ_b, and ω_T with $loss$
  $iteration \leftarrow iteration + 1$
end while

is progressively added:

$$q(\omega_t|\omega_{t-1}) = \mathcal{N}\left(\sqrt{1-\beta_t}\omega_{t-1}, \beta_t \mathbf{I}\right). \quad (6)$$

Reverse Diffusion: In the reverse diffusion process, which reconstructs ω_0 from Gaussian noise $\omega_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, MuNet [52] is used to steer the generated music towards the given condition $\mathcal{C}$, which consists of musical attributes (beat b and chord c) and text τ ($\mathcal{C} := \{\tau, b, c\}$). This is realized through the Music-Domain-Knowledge-Informed UNet (MuNet) denoiser. After the Chord Encoder $\mathbf{CE}$ and Beat Encoder $\mathbf{BE}$ encode the chord and beat information, respectively, MuNet takes the chord embedding, beat embedding, the encoded text from the Text Encoder $\mathbf{TE}$, and the output from the previous step ω_{t+1} to generate the current step's output ω_t:

$$\omega_t = \text{MuNet}\left(\mathbf{CE}(\text{chord}), \mathbf{BE}(\text{beat}), \mathbf{TE}(\text{text}), \omega_{t+1}\right). \quad (7)$$

Algorithm 3: SharpnessLoss.

Input: Logits $logits$, Targets $targets$, Sharpness coefficient α
Output: Loss value $loss$

$ce\_loss = \text{CrossEntropy}(logits, targets)$
$sharpness\_penalty = logits[targets]$
$loss = ce\_loss - \alpha \cdot \text{mean}(sharpness\_penalty)$
return $loss$

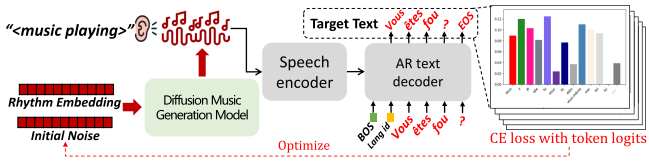


Fig. 9. Overview of our music-based attack on ST.

Fig. 9 presents the framework of our music-based attack scheme. Similar to Section IV-A, our goal is to align the translation results obtained by the autoregressive prediction model with the target sentences. However, unlike the previous approach, here we focus on optimizing the control inputs for music generation, specifically the inputs to (7).

Attack with Diffusion-based Music Generation (DMG): During the audio generation phase, *i.e.*, the denoising process of the LDM, we set the initial noise z_T as the optimization target. This noise is optimized through gradient backpropagation to ensure that the final denoised music contains adversarial elements. To enhance the effectiveness of the adversarial attack, we also include rhythm (beat and chord) in the optimization target set. We optimize the parameters θ_c of the Chord Encoder $\mathbf{CE}$ and θ_b of the Beat Encoder $\mathbf{BE}$ to refine the fundamental music properties. The goal is to make the audio generated by the LDM adversarial, ensuring it translates into specific semantic content. Detailed algorithm is presented in Algorithm 2.

Since music is not part of the typical data distribution for speech translation models, the model tends to interpret adversarial music as target semantics with lower confidence during the optimization process. This can compromise the stability of the optimization and the cross-lingual generalization of adversarial samples. To address this challenge, we employ **SharpnessLoss** (see Algorithm 3) as a replacement for Cross-Entropy Loss in this context. Specifically, we enhance the standard cross-entropy loss by optimizing the logits of the translation results. The objective is to ensure that each step of the translation process has a high probability of generating the target text, thereby sharpening the predicted distribution at each step of the autoregressive prediction process.

Enhance with Simulated Over-the-air Process: As outlined in Section III-A, a more potent attack strategy involves transmitting adversarial music through an over-the-air channel, as depicted in Fig. 4. To ensure over-the-air robustness, we simulate air transmission distortions and environmental noise by overlaying

TABLE I
SETTINGS OF SEEN AND UNSEEN LANGUAGES

Model	Attack Lang.	Test Lang. (Seen)	Test Lang. (Unseen)
Seamless	EN	EN	ZH, DE, FR, IT, ES
	EN + ZH	EN, ZH	DE, FR, IT, ES
	EN + ZH + DE	EN, ZH, DE	FR, IT, ES
	EN + ZH + DE + FR	EN, ZH, DE, FR	IT, ES
Canary	EN	EN	FR, DE, ES
	EN + FR	EN, FR	DE, ES
	EN + FR + DE	EN, FR, DE	ES
	EN + FR + DE + ES	EN, FR, DE, ES	None

Note: EN=English, ZH=Mandarin, DE=German, FR=French, IT=Italian, ES=Spanish

speech from the Librispeech dataset [54] and applying random impulse responses from [55] for reverberation. Small random noise is also added. Details are provided in Sec. SUPP-G in supplementary Appendix.

V. EVALUATIONS

A. Experimental Setup

Target Models: To thoroughly investigate the vulnerability of ST models to adversarial attacks, we selected two types of models: Standard End-to-End ST Model and Speech-to-any ST Model. Canary [9] and Seamless [7] were chosen as the state-of-the-art representatives of each category. Additionally, as Seamless is currently one of the most advanced speech translation models, we conducted extensive experiments on its various versions (Large, Expressive, M4Tv2, and Medium), which differ in model architecture and parameter sizes.

Languages: We selected different target language sets as the optimization targets during the attack and conducted tests across multiple target languages. These target languages included both **Seen** languages (encountered during adversarial optimization) and **Unseen** languages, allowing for a comprehensive evaluation of semantic attack effectiveness.

For Seamless [7], as shown in Table I, we selected *English(EN)*, *Mandarin(ZH)*, *German(DE)*, *French(FR)*, *Italian(IT)*, and *Spanish(ES)* as the target test languages, since this set represents the largest intersection of languages supported by all Seamless models. This selection ensures consistency across the different model versions. For the attacks, we tested four language sets: *EN*, *EN+ZH*, *EN+ZH+DE*, and *EN+ZH+DE+FR*, to evaluate how the number of **Seen** languages influence the attack's performance.

For Canary [9], which only supports *English*, *French*, *German*, and *Spanish*, we set up four attack combinations: *EN*, *EN+FR*, *EN+FR+DE*, and *EN+FR+DE+ES*. These four languages were also used as the target test languages.

Target semantics: We conduct experiments with 5 target semantics: "You make me sick.","Shame on you.", "Are you insane?", "Who do you think you're talking to?", and "Don't waste my time anymore.", representing the malicious semantics that attackers may inject in speech translation scenarios.

Carrier Speech Set: For the perturbation carriers in the perturbation-based attack, we select one speech sample from two distinct speakers for each of the following six languages: English (from VCTK [56]), Mandarin (from AISHELL [57]), German, French, Italian (from CommonVoice [58]), and Spanish (from VoxPopuli [59]). Samples are available in the project homepage. This setup yields 60 test cases for each attack configuration (attack method, target language, target semantics).

Attack Method: We explore the vulnerability of speech translation models using two different strategies: Perturbation-based Attack and Music-based Attack. For the perturbation-based attack, as outlined in Section IV-A, we applied adversarial perturbations to the carrier speech through gradient optimization. For the music-based attack, as described in Section IV-B, we introduced a novel adversarial music optimization scheme that leverages diffusion-based music generation. This approach is more covert, because it imitates background music and environmental noise that are not easily noticeable.

Evaluation Metrics: We used a variety of metrics to comprehensively evaluate the two perspectives of the attack: adversarial audio quality and attack effectiveness.

To evaluate the quality of adversarial speech, we utilize three objective metrics: Perceptual Evaluation of Speech Quality (**PESQ**) [60], Speaker Vector Cosine Similarity (**VSIM**) [61], and Speaker Vector Cosine Similarity specific to Seamless (**VSIM − E**) [7], along with a subjective metric, Mean Opinion Score(**MOS**). In detail, PESQ assesses speech quality (*i.e.*, imperceptibility) by taking into account the nuances of the human auditory system. VSIM measures speaker similarity to evaluate fidelity, with higher values indicating greater similarity. Following prior works [61], [62], we compute VSIM using the speaker encoder from the Resemblyzer package [63]. For VSIM-E, we use the speaker encoder of Seamless [7].

To evaluate the effectiveness of adversarial attack schemes, *i.e.*, the similarity between the model output and the attack target, we employed two metrics. Traditional metrics like Word Error Rate (WER) are insufficient here, as they penalize semantically similar but lexically different translations. For instance, "shame on you" and "you should be ashamed of yourself" would yield a high WER despite near-identical meanings. The first metric is the semantic similarity between the translation output and the target, measured by the embedding similarity (**ESIM**) from a pre-trained BERT model [64]. This metric is widely used in translation evaluations [15], [65]. The second metric is **NSCORE**, which assesses the semantic entailment relationship between the translation result and the target text, following the approach in Natural Language Inference (NLI) tasks [66], [67].

To establish appropriate semantic similarity thresholds for measuring attack Success Rate (**SR**), we leveraged similarity scores, which typically yield very low values between semantically unrelated sentences (examples shown in Table II. For each target meaning, we used ChatGPT-4 to generate six semantically equivalent but structurally diverse paraphrases (e.g., "shame on you" vs. "you should be ashamed of yourself"). We then calculated **ESIM** and **NSCORE** values between the original text and these variations to obtain the lowest similarity scores, denoted as Γ_e and Γ_n, which serve as thresholds to determine semantic consistency between target semantic and model output.

TABLE II
EXAMPLES OF ESIM SCORES BETWEEN SEMANTICALLY RELATED(SUCCESSFUL) AND UNRELATED(UNSUCCESSFUL) TEXT PAIRS

Target semantic	Adversarial output	ESIM(↑)	NSCORE(↑)
Shame on you.	Have you no shame?	**0.5644**	**0.6581**
	The bus is running late today.	0.0492	0.0077
You make me sick.	You revolt me.	**0.7134**	**0.4215**
	We need to buy more coffee.	0.0942	0.0072

Examples and prompts for paraphrase generation are shown in Fig. 13 in Supplementary Resources.

B. Perturbation-Based Attack

In this section, we first conduct a detailed analysis of the attack effectiveness of the generated perturbations, followed by an evaluation of their perceptual quality.

1) Attack Effectiveness: We begin by assessing the attack's effectiveness on the default model, Seamless Large, before extending our analysis to other models.

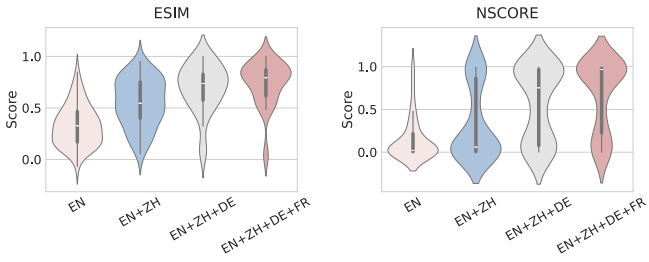
We begin with preliminary experiments in a single-language attack scenario, specifically investigating the effectiveness of targeted adversarial attacks in a translation task from language A to language B. The experimental results, shown in Table III, evaluate the impact of attacks under varying perturbation levels. The results demonstrate that adversarial attacks can be effectively applied to any target language. Notably, we observed that the attack's effectiveness is closely related to the target language but shows minimal dependence on the source language. This phenomenon arises because LLM models like Seamless employ a paradigm that maps input languages into a language-agnostic semantic space, eliminating the need to specify the source language token during translation. Therefore, in subsequent experiments, we focus on analyzing scenarios involving different target languages.

2) Enhancement Based on More Seen Languages: As briefly mentioned in Section IV-A, introducing more Seen languages during the generation of adversarial perturbations enhances cross-language generalization.

As shown in Table IV, increasing the number of Seen languages enhances the attack transferability to Unseen languages. This indicates that multilingual translation models exhibit semantic alignment across different languages, and optimizing perturbations with more Seen languages results in perturbations that more closely align with the target semantics, as shown in Fig. 6. More results under different attack intensities refer to Tabs. XVII, XVIII in supplementary Appendix.

In Fig. 10, we illustrate the data distributions of ESIM and NSCORE, using Spanish as the target Unseen language, under the scenario of increasing the number of Seen languages. Combining the insights from Table IV and Fig. 10, we observe that incorporating more attack languages improves the generalizability of adversarial perturbations to Unseen languages.

3) Enhancement Based on Target Cycle Optimization: As described in Section IV-A and Fig. 7, we can perform a Target Cycle Optimization(TCO) on the attack targets to generate semantically similar targets that are easier to attack. We tested this approach on the default target model, Seamless Large, and



Note: EN=English, ZH=Mandarin, DE=German, FR=French, IT=Italian, ES=Spanish

Fig. 10. Standard violin plot showing the distribution of ESIM and NSCORE scores for Spanish (Unseen) translation under varying numbers of attack languages ($\epsilon = 0.1$). Note: EN=English, ZH=Mandarin, DE=German, FR=French, IT=Italian, ES=Spanish.

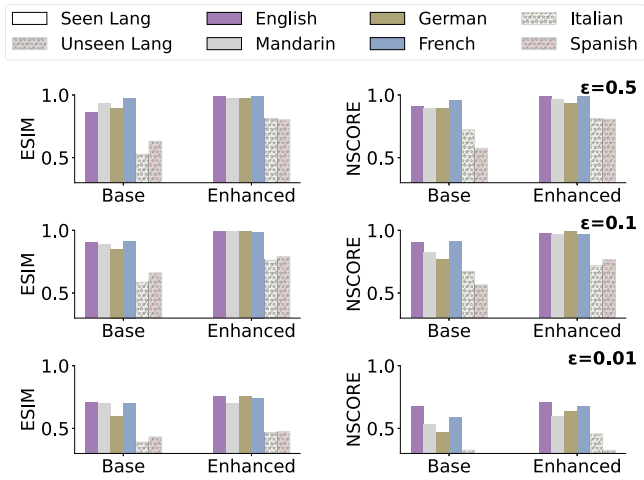


Fig. 11. The attack effectiveness of adversarial perturbations before and after TCO enhancement. TCO significantly enhances attack effectiveness, particularly for Unseen languages, i.e., Italian and Spanish.

the results are shown in Fig. 11. Under different perturbation intensities, the effectiveness of adversarial attacks consistently improves after TCO application. This improvement is particularly notable in the transferability to Unseen languages, which significantly outperforms the model before enhancement. This is because the target text generated through TCO is more compatible with different languages and is closer to the central semantics for the target model. Updated target examples are provided in supplementary Appendix Tab. XV.

4) Perceptual Evaluation: To more comprehensively explore the effects of perturbation attacks, we applied different range constraints to the perturbations, as described in Section IV-A. Larger perturbation ranges are easier to perceive but tend to yield better attack results. In this study, we investigated perturbation ranges set to $\epsilon = 0.5$, 0.1, and 0.01. Table V presents the perceptual metrics for adversarially perturbed audio when the target model is Seamless Large. We optimized the perturbations using different numbers of attack languages as targets and compared the results with random perturbations of the same magnitude applied to the original speech.

TABLE III
EFFECTIVENESS OF ADVERSARIAL ATTACKS AT DIFFERENT PERTURBATION LEVELS IN SINGLE-LANGUAGE TRANSLATION SCENARIOS, MEASURED BY AVERAGE ESIM, NSCORE AND SR. "SRC" DENOTES THE SOURCE LANGUAGE, WHILE "TGT" REFERS TO THE TARGET LANGUAGE.

ϵ	Src \ Tgt	EN			ZH			DE			FR			IT			ES		
		ESIM(↑)	NSCORE(↑)	SR(↑)	ESIM(↑)	NSCORE(↑)	SR(↑)	ESIM(↑)	NSCORE(↑)	SR(↑)	ESIM(↑)	NSCORE(↑)	SR(↑)	ESIM(↑)	NSCORE(↑)	SR(↑)	ESIM(↑)	NSCORE(↑)	SR(↑)
0.5	EN	0.9449	0.8869	10/10	0.9791	0.8100	10/10	0.8705	0.7105	9/10	0.9987	0.9850	10/10	0.8078	0.5928	9/10	0.9359	0.8827	9/10
	ZH	0.9951	0.9844	10/10	0.9358	0.7873	10/10	0.8749	0.8098	9/10	1.0000	0.9851	10/10	0.9155	0.8850	9/10	0.9416	0.8855	9/10
	DE	1.0000	0.9846	10/10	0.9648	0.8817	10/10	0.8643	0.7881	8/10	0.8952	0.8008	10/10	0.7476	0.5949	7/10	1.0000	0.9848	10/10
	FR	0.8455	0.6939	9/10	1.0000	0.9859	10/10	0.9346	0.7927	9/10	0.8651	0.8794	9/10	0.9621	0.9200	9/10	0.9808	0.8880	10/10
	IT	0.9827	0.9842	10/10	0.9536	0.9616	10/10	0.7374	0.6343	9/10	0.9586	0.9474	10/10	0.9571	0.8038	10/10	0.8970	0.8603	10/10
	ES	0.9987	0.9846	10/10	0.9400	0.7908	9/10	0.7621	0.6094	8/10	0.7602	0.7320	8/10	0.9629	0.8843	10/10	0.9023	0.8080	10/10
0.1	EN	0.8574	0.7137	8/10	0.9552	0.8915	9/10	0.7748	0.7261	8/10	0.9321	0.8861	9/10	0.9047	0.8192	9/10	0.9006	0.9768	10/10
	ZH	1.0000	0.9846	10/10	0.8809	0.7884	9/10	0.8042	0.6101	9/10	0.8441	0.8627	9/10	0.9360	0.8192	10/10	0.8341	0.7908	8/10
	DE	0.8229	0.5927	9/10	0.8387	0.8253	9/10	0.5300	0.3321	5/10	0.9679	0.9802	10/10	0.7321	0.5960	8/10	0.8626	0.7182	8/10
	FR	0.8579	0.6916	8/10	0.9591	0.8860	10/10	0.7057	0.5094	6/10	0.8588	0.7350	9/10	0.8954	0.8228	10/10	0.6614	0.4015	5/10
	IT	0.9373	0.8967	9/10	0.9490	0.8879	10/10	0.7960	0.6590	8/10	0.9302	0.9653	10/10	0.9118	0.8858	9/10	0.9461	0.9249	10/10
	ES	0.9084	0.8334	8/10	0.8570	0.7996	7/10	0.6418	0.4208	5/10	0.7899	0.6153	10/10	0.8959	0.8514	9/10	0.6179	0.4348	5/10
0.01	EN	0.7672	0.5545	7/10	0.7832	0.6041	7/10	0.5118	0.1111	5/10	0.6584	0.6014	6/10	0.6312	0.5751	7/10	0.6114	0.3465	5/10
	ZH	0.6899	0.6057	6/10	0.5134	0.3060	5/10	0.5136	0.2168	6/10	0.9188	0.7878	10/10	0.6919	0.5069	6/10	0.9270	0.8465	10/10
	DE	0.4083	0.2915	5/10	0.7132	0.5945	7/10	0.3274	0.2452	5/10	0.6078	0.5147	7/10	0.5650	0.4992	6/10	0.3720	0.1101	4/10
	FR	0.3923	0.2326	2/10	0.6216	0.3099	7/10	0.5256	0.4209	5/10	0.7566	0.6969	8/10	0.3142	0.2258	3/10	0.4552	0.4745	5/10
	IT	0.5082	0.4083	6/10	0.3361	0.2649	3/10	0.2327	0.1717	5/10	0.5172	0.3997	6/10	0.3502	0.2024	4/10	0.5275	0.4367	6/10
	ES	0.5362	0.4320	6/10	0.2591	0.1814	2/10	0.2413	0.1721	4/10	0.3383	0.2292	5/10	0.4166	0.2470	4/10	0.4250	0.2633	4/10

Note: EN=English, ZH=Mandarin, DE=German, FR=French, IT=Italian, ES=Spanish

TABLE IV
ATTACK EFFECTIVENESS OF ADVERSARIAL PERTURBATIONS ($\epsilon = 0.1$), MEASURED BY AVERAGE ESIM, NSCORE, AND SR. BLUE-HIGHLIGHTED AREAS INDICATE TESTS CONDUCTED ON **SEEN** LANGUAGES.

Attack with	Target	Similarity with Target		SR(↑)
		ESIM(↑)	NSCORE(↑)	
English	English	0.8973	0.7855	52/60
	Mandarin	0.3900	0.1717	19/60
	German	0.2999	0.1214	27/60
	French	0.3645	0.1794	30/60
	Italian	0.3148	0.1515	14/60
	Spanish	0.3275	0.1467	22/60
English Mandarin	English	0.9234	0.9221	58/60
	Mandarin	0.9844	0.9677	59/60
	German	0.5823	0.4216	37/60
	French	0.6036	0.3901	39/60
	Italian	0.4854	0.3459	34/60
	Spanish	0.5492	0.3254	35/60
English Mandarin German	English	0.9290	0.9010	58/60
	Mandarin	0.9479	0.8877	56/60
	German	0.8771	0.8124	58/60
	French	0.7281	0.6866	53/60
	Italian	0.6415	0.6593	49/60
	Spanish	0.6964	0.5705	51/60
English Mandarin German French	English	0.9238	0.9015	58/60
	Mandarin	0.9222	0.8511	57/60
	German	0.8912	0.8335	56/60
	French	0.9356	0.9118	57/60
	Italian	0.7026	0.7450	53/60
	Spanish	0.7414	0.6804	55/60

TABLE V
PERCEPTION INFLUENCE OF ADVERSARIAL PERTURBATION. INDICATORS MARKED WITH "$*$" CORRESPOND TO RANDOM PERTURBATIONS.

ϵ	Attack with	Adversarial perturbation			Random perturbation		
		PESQ(↑)	VSIM(↑)	VSIM-E(↑)	PESQ*(↑)	VSIM*(↑)	VSIM-E*(↑)
0.5	EN	1.1395	0.4661	0.2617	1.0658	0.4886	-0.0942
	EN+ZH	1.0692	0.4756	0.2492	1.0052	0.4881	-0.1103
	EN+ZH+DE	1.2289	0.4705	0.2413	1.0443	0.4831	-0.1124
	EN+ZH+DE+FR	1.1191	0.4724	0.2369	1.0248	0.4768	-0.1091
0.1	EN	1.4102	0.6146	0.4172	1.4541	0.5911	0.1452
	EN+ZH	1.4077	0.6003	0.4037	1.4050	0.5746	0.1252
	EN+ZH+DE	1.3930	0.5915	0.3942	1.3687	0.5581	0.1096
	EN+ZH+DE+FR	1.3811	0.5881	0.3917	1.3579	0.5592	0.1049
0.01	EN	2.3671	0.8346	0.6710	2.6614	0.8366	0.5107
	EN+ZH	2.3297	0.8286	0.6654	2.6300	0.8309	0.5009
	EN+ZH+DE	2.3089	0.8218	0.6600	2.6063	0.8259	0.4935
	EN+ZH+DE+FR	2.3045	0.8228	0.6577	2.6093	0.8257	0.4959

Note: EN=English, ZH=Mandarin, DE=German, FR=French

TABLE VI
GENERALIZABILITY EVALUATION OF ADVERSARIAL PERTURBATIONS ACROSS MULTIPLE MODELS, WITH ENGLISH USED AS THE ATTACK AND TARGET LANGUAGE

ϵ	Target model	Similarity with Original		Similarity with Target		SR(↑)
		ESIM(↓)	NSCORE(↓)	ESIM(↑)	NSCORE(↑)	
0.5	Seamless Large	0.0285	0.0288	0.9612	0.9198	59/60
	Seamless Medium	0.0332	0.0291	0.9927	0.9635	60/60
	Seamless M4Tv2	0.0383	0.0254	0.9201	0.8980	57/60
	Seamless Expressive	0.0452	0.0374	0.8925	0.8224	55/60
0.1	Seamless Large	0.0317	0.0303	0.8973	0.7855	52/60
	Seamless Medium	0.0422	0.0343	0.9240	0.8780	57/60
	Seamless M4Tv2	0.0412	0.0272	0.9022	0.8375	54/60
	Seamless Expressive	0.0813	0.0386	0.7815	0.6840	49/60
0.01	Seamless Large	0.2154	0.0921	0.5503	0.4207	32/60
	Seamless Medium	0.1223	0.0748	0.7514	0.6205	44/60
	Seamless M4Tv2	0.2449	0.1318	0.4980	0.3361	25/60
	Seamless Expressive	0.2099	0.0996	0.6003	0.5031	36/60

The results show that adversarial perturbations exhibit better perceptual quality than random perturbations of the same magnitude, particularly in maintaining speaker timbre and acoustic environment (VSIM-E). This is because our perturbations are specifically designed to avoid high or low frequency bands, as explained in Section IV-A. This approach significantly minimizes the impact on the core content of the speech (PESQ) and preserves speech style (VSIM, VSIM-E). A more detailed analysis and discussion of the perceptual quality impact of perturbations are provided in Appendix Sec. SUPP-D.

5) Generalizability: We evaluated the effectiveness of the proposed method across different models. We conducted extensive tests on multiple models to assess the generalizability of the adversarial approach. As illustrated in Table VI, all examples and target semantics were tested with English serving as both the attack and target language. Crucially, the translated semantics of the audio (ESIM, NSCORE) post-attack closely align with the target semantics while exhibiting a significant deviation from the original semantics.

Additionally, to further investigate the generalization capability of the proposed method, we introduced an additional model, Canary [9], which is not part of the Seamless model family, for testing. We conducted attacks on Canary using different numbers of languages, detailed in Tab. XX in SUPP. The proposed method demonstrates strong attack performance on the model outside the Seamless model family. Furthermore, the enhancement effect with more Seen languages remains consistent with the findings of previous experiments. Combined with the results in Table VI, these findings demonstrate that the proposed perturbation-base attack method effectively performs attacks across different models.

6) Voice Content Protection: The adversarial noise is not only instrumental in achieving the desired target semantics

TABLE VII
EVALUATION OF THE ORIGINAL SEMANTIC HIDING CAPABILITY FOR ADVERSARIAL PERTURBATIONS AND RANDOM PERTURBATIONS ACROSS MULTIPLE MODELS, USING ENGLISH AS BOTH THE ATTACK AND TARGET LANGUAGE

ϵ	Target model	Adversarial perturbation		Random perturbation	
		ESIM(↓)	NSCORE(↓)	ESIM(↓)	NSCORE(↓)
0.5	Seamless large	0.0285	0.0288	0.0633	0.0583
	Seamless medium	0.0332	0.0291	0.0969	0.0724
	Seamless m4tv2	0.0383	0.0254	0.1066	0.0848
	Seamless expressive	0.0452	0.0374	0.0697	0.2294
0.1	Seamless large	0.0317	0.0303	0.6185	0.4391
	Seamless medium	0.0422	0.0343	0.5999	0.4307
	Seamless m4tv2	0.0412	0.0272	0.7537	0.5129
	Seamless expressive	0.0813	0.0386	0.6793	0.5126
0.01	Seamless large	0.2154	0.0921	0.9063	0.7732
	Seamless medium	0.1223	0.0748	0.8975	0.7854
	Seamless m4tv2	0.2449	0.1318	0.9650	0.8477
	Seamless expressive	0.2099	0.0996	0.9319	0.7596

control but also possesses the potential to severely disrupt the comprehension of the original content, thereby serving as a viable content protection measure. As illustrated in Table VI, the translated semantics of the perturbed audio (quantified by ESIM and NSCORE) closely align with the target semantics while exhibiting a significant deviation from the original semantics. Notably, even a very low perturbation magnitude ($\epsilon = 0.01$) leads to a significant degradation of the original semantics. This evidence strongly indicates the feasibility of robust voice content protection.

Furthermore, we investigated the impact of random perturbations of equivalent magnitude on the original semantics. As shown in Table VII, under the same magnitude constraint, the adversarial perturbation causes the translated semantics to deviate significantly further from the original semantics, effectively achieving semantic hiding. Conversely, the semantic hiding capability of random perturbation is substantially weaker than that of its adversarial counterpart, particularly at low perturbation magnitudes ($\epsilon = 0.1$ and $\epsilon = 0.01$), where it causes negligible interference to the original semantics.

C. Music-Based Attack

As described in Section IV-B, we also explored the method of attacking using adversarial music.

The Seamless model family does not require source language specification, as it maps input audio directly to a multilingual semantic space via speech understanding. Similarly, preliminary analysis of the Canary model indicates that the source language has limited impact on translation outputs. Accordingly, we set English as the default source language for all translation models during evaluation.

1) Attack Effectiveness: To further investigate the impact of adversarial music, we expanded the target semantics based on the original set.[1] Table VIII presents the attack performance when targeting six different languages. The results demonstrate effectiveness comparable to the perturbation outcomes reported in Table III.

[1]The newly added target semantics are: "This is unbelievable.", "I can't stand you.", "This is ridiculous.", "Stop bothering me.", and "What's wrong with you?"

TABLE VIII
ATTACK ABILITY OF ADVERSARIAL MUSIC IN SINGLE-LANGUAGE TRANSLATION SCENARIOS

Target Language	ESIM(↑)	NSCORE(↑)	SR(↑)
English	0.7879	0.7507	9/10
Mandarin	0.9669	0.9314	10/10
German	0.7281	0.6139	8/10
French	0.7783	0.7646	8/10
Italian	0.6216	0.4871	6/10
Spanish	0.8491	0.8823	9/10

TABLE IX
ATTACK ABILITY OF ADVERSARIAL MUSIC WITH SEAMLESS LARGE AS TARGET MODEL. BLUE-HIGHLIGHTED AREAS INDICATE TESTS CONDUCTED ON **SEEN** LANGUAGES.

Attack with	Target	Similarity With Target		SR(↑)
		ESIM(↑)	NSCORE(↑)	
English	English	0.7879	0.7507	9/10
	Mandarin	0.5152	0.4257	6/10
	German	0.5706	0.4236	6/10
	French	0.4643	0.5759	7/10
	Italian	0.4877	0.6616	7/10
	Spanish	0.4408	0.4661	4/10
English Mandarin	English	0.8434	0.7893	9/10
	Mandarin	0.8362	0.6615	10/10
	German	0.7633	0.6370	9/10
	French	0.6396	0.5117	8/10
	Italian	0.6199	0.6236	7/10
	Spanish	0.6691	0.5366	7/10
English Mandarin German	English	0.8493	0.8823	9/10
	Mandarin	0.8516	0.8559	9/10
	German	0.9466	0.7901	10/10
	French	0.6953	0.6626	8/10
	Italian	0.7277	0.7611	8/10
	Spanish	0.7412	0.6852	8/10
English Mandarin German French	English	0.9267	0.9821	10/10
	Mandarin	0.8899	0.8915	9/10
	German	0.8519	0.8866	9/10
	French	0.8804	0.8851	9/10
	Italian	0.7434	0.8818	9/10
	Spanish	0.8021	0.8704	10/10

The adversarial music generation process optimizes only the initial latent code and rhythm encoding during the diffusion process. To ensure experimental control, a fixed prompt was used as the text-to-music input. An exploration of different music generation prompts is detailed in Sec. SUPP-C.

2) Enhancement Based on More Seen Languages: Table IX show the results of attacks using different Seen languages. We observe that: (1) the generated adversarial music demonstrates strong attack capabilities on seen languages; (2) as the number of Seen languages increases, the adversarial music exhibits better generalization across multilingual scenarios; (3) overall, the adversarial music effectively attacks the target model.

3) Enhancement Based on Target Cycle Optimization: As outlined in Fig. 7, Target Cycle Optimization (TCO) can be applied to the attack targets, generating semantically similar targets that are more susceptible to attack. Similar to the experiments discussed in Section V-B3, we tested this approach

TABLE X
EVALUATION OF THE IMPACT OF DIFFERENT AUDIO SIGNAL PROCESSING TECHNIQUES ON THE EFFECTIVENESS OF ADVERSARIAL PERTURBATIONS AND ADVERSARIAL MUSIC

Type		Perturbation						Music					
Processing		None	LPF	MP3	Quant	Noise	Resample	None	LPF	MP3	Quant	Noise	Resample
Similarity With Original	ESIM(↓)	**0.0317**	0.0932	0.0879	0.1367	0.0487	0.1037	-	-	-	-	-	-
	NSCORE(↓)	**0.0303**	0.0345	0.0478	0.1166	0.0355	0.0556	-	-	-	-	-	-
Similarity With Target	ESIM(↑)	**0.8973**	0.5803	0.5361	0.2675	0.7490	0.4175	**0.7879**	0.7196	0.6378	0.4175	0.5845	0.4638
	NSCORE(↑)	**0.7855**	0.3872	0.3349	0.1345	0.5767	0.1723	**0.7507**	0.6585	0.6125	0.3704	0.5577	0.3034
SR(↑)		**52/60**	35/60	27/60	12/60	45/60	21/60	**9/10**	8/10	8/10	4/10	7/10	4/10

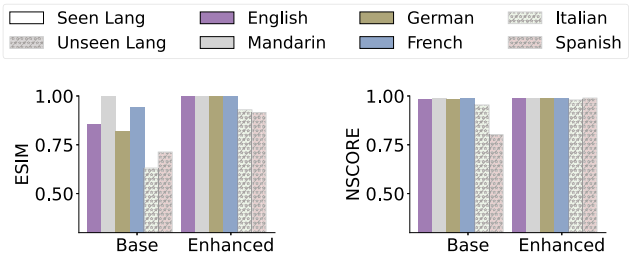


Fig. 12. The attack effectiveness of adversarial music is enhanced through Target Cycle Optimization.

on the default target model, Seamless Large, and the results are presented in Fig. 12. The application of TCO significantly improves the effectiveness of adversarial attacks, as indicated by higher semantic similarity between the translated results and the target. The improvement is especially evident in the transferability to Unseen languages, where attack performance improve significantly after applying TCO. The enhanced targets generated through TCO are better aligned with various languages and are closer to the central semantics in the target model. Updated target examples are provided in Supplementary Table XV.

4) Generalizability: We conducted additional experiments on Canary [9], and the results are shown in Table XIII. Consistent with the results in Table XII, this further demonstrates the generalization capability of adversarial music. Furthermore, the enhancement effect with more Seen languages remains consistent with the findings of previous experiments.

5) Physical Test: As described in Section III-A, a more severe attack method involves transmitting adversarial music over the air, as illustrated in Fig. 4. To implement this, we integrated simulated air-channel transmission distortions into the adversarial perturbation optimization process. Details of these distortions are provided in supplementary Sec. SUPP-G.

We evaluated adversarial music attacks on two models: Seamless Large and Canary. A consumer-grade speaker was used for playback, while a consumer-grade microphone and a smartphone captured the audio to simulate typical over-the-air conditions. The specifications of the devices are detailed in Fig. 17 in Supplementary Resources.[2]

For each attack, six adversarial music samples were generated and tested multiple times to ensure stability, resulting in 60 test samples per target language. The results, summarized in Table XIV, indicate that adversarial music achieves an attack success rate of approximately 50% across various models and devices in over-the-air attack scenarios. These findings suggest that adversarial music could be exploited to inject malicious semantics into real-time speech translation conferences or conversations, posing significant security risks.

D. User Study

In addition to the attack effectiveness and objective quality evaluations, we also conducted subjective experimental assessments on both adversarial perturbations and adversarial music. In this test, 20 participants were invited to rate the quality of speech overlaid with adversarial perturbations and the generated adversarial music.[3] To serve as a baseline, random white noise matching the energy intensity of each adversarial perturbation was generated. Similarly, white noise with equivalent energy intensity was created for each piece of adversarial music. The detailed scoring criteria and the statistical results for both perturbations and music can be found in the Supplementary Resources, specifically in Table XIX, Figs. 15 and 16.

As the perturbation strength increases, the quality scores demonstrate a downward trend. However, adversarial perturbations consistently demonstrate better perceptual quality compared to random perturbations of the same strength, particularly at higher perturbation levels. With the default $\epsilon = 0.1$, our results indicate that most adversarial perturbations do not significantly affect the perception of speech content.

For the generated adversarial music, our evaluation shows that adversarial music demonstrates better perceptual quality compared to random perturbations of the same strength. Furthermore, the generated music receives high scores, demonstrating the imperceptibility of our adversarial music approach.

VI. POSSIBLE DEFENSES

To evaluate potential countermeasures against the identified security vulnerability, we conducted a series of defense experiments targeting the proposed adversarial perturbations and music-based attacks. Specifically, various audio signal processing techniques were applied to introduce distortions, aiming to disrupt the adversarial effectiveness of proposed attacks. These techniques included filtering (6 kHz low-pass, denoted as LPF), compression (64 kbps, MP3), noise addition (SNR 64 dB, Noise), quantization (8-bit, Quant), and resampling (12 kHz, Resample). The results are presented in Table X.

[2]Experiments were conducted in a room measuring 4.37 m × 2.35 m × 2.95 m, with the microphone and speaker placed 50 cm apart.

[3]Participants (aged 19–31) included university students, professors, and non-academic professionals, spanning both computer science and non-computer-related educational backgrounds.

TABLE XI
EVALUATION OF THE IMPACT OF VARIOUS AUDIO SIGNAL PROCESSING TECHNIQUES ON SPEECH PERCEPTUAL QUALITY

Processing		None	LPF	MP3	Quant	Noise	Resample
Similarity With Original	PESQ(↑)	1.4102	1.4102	1.4100	1.4104	1.4102	1.4100
	VSIM(↑)	0.6146	0.6124	0.6152	0.6136	0.6146	0.6133
	VSIM-E(↑)	0.4172	0.3895	0.3954	0.3628	0.4086	0.3611
Similarity With Adversarial	PESQ(↑)	4.5000	4.4995	4.3787	4.0504	4.4943	4.4994
	VSIM(↑)	1.0000	0.9992	0.9991	0.9982	1.0000	0.9995
	VSIM-E(↑)	1.0000	0.9722	0.9739	0.9355	0.9978	0.9407

TABLE XII
GENERALIZABILITY EVALUATION OF ADVERSARIAL MUSIC ACROSS MULTIPLE MODELS, WITH ENGLISH USED AS THE ATTACK AND TARGET LANGUAGE

Target model	Similarity With Target		SR(↑)
	ESIM(↑)	NSCORE(↑)	
Seamless Large	0.7879	0.7507	9/10
Seamless Medium	0.9211	0.8100	9/10
Seamless M4tv2	0.8017	0.8164	10/10
Seamless Expressive	0.9142	0.9595	10/10

TABLE XIII
ATTACK ABILITY OF ADVERSARIAL MUSIC WITH CANARY AS TARGET MODEL. BLUE-HIGHLIGHTED AREAS INDICATE TESTS CONDUCTED ON **SEEN** LANGUAGES

Attack with	Target	Similarity With Target		SR(↑)
		ESIM(↑)	NSCORE(↑)	
English	English	0.7899	0.7588	7/10
	French	0.5881	0.3989	6/10
	German	0.5863	0.3620	9/10
	Spanish	0.5661	0.4407	7/10
English French	English	0.9817	0.9543	10/10
	French	0.9397	0.9117	9/10
	German	0.7567	0.7013	9/10
	Spanish	0.6729	0.6010	7/10
English French German	English	0.9616	0.9730	10/10
	French	1.0000	0.9862	10/10
	German	0.9984	0.9865	10/10
	Spanish	0.8544	0.8846	9/10
English	English	0.9935	0.9877	10/10
French	French	0.9988	0.9862	10/10
German	German	0.9247	0.8242	10/10
Spanish	Spanish	0.9800	0.9856	10/10

The experimental results indicate that adversarial perturbations and adversarial music exhibit a certain degree of robustness to audio processing. However, certain techniques, particularly quantization and resampling, can significantly impact the attack effectiveness. This finding suggests that, in the absence of cost concerns, resisting adversarial audio attacks is feasible. However, based on the results from perturbation removal experiments, while these processing techniques mitigate the intensity of adversarial attacks, they do not fully restore the semantic integrity of the original speech. Moreover, these methods may interfere with the semantic information of the original audio, thereby reducing its usability.

Furthermore, we analyzed the impact of pre-processing on speech quality. As shown in Table XI, by observing the similarity between the pure adversarial speech (processing is None) and the adversarial speech after various processing techniques, and the original speech, we find that processing further degrades speech quality, especially Quant and Resample. A similar phenomenon is observed when examining the similarity between the processed speech and the adversarial speech.

TABLE XIV
ATTACK ABILITY OF ADVERSARIAL MUSIC IN OVER-THE-AIR ENGLISH-TARGETED TRANSLATION SCENARIOS

Target model	Device	Target Language	SR(↑)
Seamless large	Microphone	English	31/60
		Mandarin	34/60
		German	38/60
		French	33/60
		Italian	29/60
		Spanish	27/60
	Cell Phone	English	28/60
		Mandarin	35/60
		German	38/60
		French	34/60
		Italian	33/60
		Spanish	25/60
Canary	Microphone	English	27/60
		French	47/60
		German	30/60
		Spanish	36/60
	Cell Phone	English	29/60
		French	38/60
		German	34/60
		Spanish	41/60

VII. CONCLUSION

In this paper, we explored the vulnerability of ST systems to targeted adversarial attacks and proposed two strategies: a perturbation-based attack and an innovative music-generation-based approach. We introduced several methods to enhance adversarial attacks on ST models, including Multi-language Enhancement and Target Cycle Optimization. Extensive experiments were conducted using various source and target language pairs, demonstrating the susceptibility of current ST systems to adversarial attacks. We hope our research raises awareness of the security challenges in ST systems and contributes to efforts to strengthen their robustness.

VIII. ETHICAL CONSIDERATIONS

This research investigates adversarial attacks on speech translation systems, specifically identifying vulnerabilities in state-of-the-art models. While we believe our work contributes positively to enhancing system security, we strictly adhered to key ethical principles to minimize potential risks:

- *Public Systems:* Our study focused exclusively on publicly available models, with the goal of informing the academic community and developers about potential security threats.
- *User Impact and Privacy:* For experiments involving subjective audio quality evaluations, participants were fully informed about the study's purpose and the security risks associated with adversarial attacks. Crucially, no personal data was collected, thereby ensuring participants' privacy and anonymity.

- *Responsible Disclosure:* While we highlight the risks of adversarial attacks, we do not endorse or encourage their use for malicious purposes. The insights derived from this research are solely intended to help improve the security and robustness of speech translation systems, ensuring their safe deployment in real-world applications.

Through this work, we aim to make a positive contribution to the field of speech translation security, assisting developers in building more resilient systems capable of withstanding adversarial manipulations.

REFERENCES

[1] S. Dhawan, "Speech to speech translation: Challenges and future," *Int. J. Comput. Appl. Technol. Res.*, vol. 11, no. 03, pp. 36–55, 2022.

[2] Y. Wang et al., "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 319–352, Firstquarter 2023.

[3] A. Lavie et al., "JANUS-III: Speech-to-speech translation in multiple languages," in *Proc. 1997 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1997, vol. 1, pp. 99–102.

[4] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin, Germany: Springer Science & Business Media, 2013.

[5] S. Nakamura et al., "The atr multilingual speech-to-speech translation system," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 365–376, Mar. 2006.

[6] H. Inaguma et al., "Espnet-st: All-in-one speech translation toolkit," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics: System Demonstrations*, 2020, pp. 302–311.

[7] L. Barrault et al., "Seamless: Multilingual expressive and streaming speech translation," 2023, *arXiv:2312.05187*.

[8] H. Wang, Z. Xue, Y. Lei, and D. Xiong, "End-to-end speech translation with mutual knowledge distillation," in *Proc. 2024 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 11306–11310.

[9] NVIDIA, "Canary," (n.d.). [Online]. Available: https://huggingface.co/nvidia/canary-1b

[10] H. Ney, "Speech translation: Coupling of recognition and translation," in *Proc. 1999 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1999, vol. 1, pp. 517–520.

[11] E. Matusov, S. Kanthak, and H. Ney, "On the integration of speech recognition and statistical machine translation," in *Proc. Interspeech*, 2005, pp. 3177–3180.

[12] A. B´erard, O. Pietquin, L. Besacier, and C. Servan, "Listen and translate: A proof of concept for end-to-end speech-to-text translation," *NIPS Workshop End-To-End Learn. Speech Audio Process.*, 2016.

[13] S. Bansal, H. Kamper, K. Livescu, A. Lopez, and S. Goldwater, "Low-resource speech-to-text translation," in *Proc. Interspeech 2018*, 2018, pp. 1298–1302.

[14] J. Iranzo-Sánchez et al., "Europarl-ST: A multilingual corpus for speech translation of parliamentary debates," in *Proc. 2020 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 8229–8233.

[15] L. Barrault et al., "SeamlessM4T-massively multilingual & multimodal machine translation," 2023, *arXiv:2308.11596*.

[16] C.-y. Huang, Y. Y. Lin, H.-y. Lee, and L.-s. Lee, "Defending your voice: Adversarial attack on voice conversion," in *Proc. 2021 IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 552–559.

[17] Z. Yu, S. Zhai, and N. Zhang, "AntiFake: Using adversarial audio to prevent unauthorized speech synthesis," in *Proc. 2023 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2023, pp. 460–474.

[18] Z. Liu, Y. Zhang, and C. Miao, "Protecting your voice from speech synthesis attacks," in *Proc. 39th Annu. Comput. Secur. Appl. Conf.*, 2023, pp. 394–408.

[19] Z. Yu, Y. Chang, N. Zhang, and C. Xiao, "Smack: Semantically meaningful adversarial audio attack," in *Proc. 32nd USENIX Secur.*Symp., 2023, pp. 3799–3816.

[20] P. Cheng et al., "ALIF: Low-cost adversarial audio attacks on black-box speech platforms using linguistic features," in *Proc. 2024 IEEE Symp. Secur. Privacy*, 2023, pp. 56–56.

[21] X. Yuan et al., "CommanderSong: A systematic approach for practical adversarial voice recognition," in *Proc. 27th USENIX Secur. Symp.*, 2018, pp. 49–64.

[22] N. Carlini et al., "Hidden voice commands," in *Proc. 25th USENIX Secur. Symp.*, 2016, pp. 513–530.

[23] H. Yakura and J. Sakuma, "Robust audio adversarial example for a physical attack," in *Proc. 28th Int. Joint Conf. Artif. Intell.. Int. Joint Conf. Artif. Intell. Org.*, 2019, pp. 5334–5341.

[24] Y. Chen et al., "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices," in *Proc. 29th USENIX Secur. Symp.*, 2020, pp. 2667–2684.

[25] R. Duan, Z. Qu, S. Zhao, L. Ding, Y. Liu, and Z. Lu, "Perception-aware attack: Creating adversarial music via reverse-engineering human perception," in *Proc. 2022 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2022, pp. 905–919.

[26] H. Liu et al., "When evil calls: Targeted adversarial voice over ip network," in *Proc. 2022 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2022, pp. 2009–2023.

[27] G. Chen et al., "Who is real Bob? Adversarial attacks on speaker recognition systems," in *Proc. 2021 IEEE Symp. Secur. Privacy*, 2021, pp. 694–711.

[28] X. Li, J. Ze, C. Yan, Y. Cheng, X. Ji, and W. Xu, "Enrollment-stage backdoor attacks on speaker recognition systems via adversarial ultrasound," *IEEE Internet Things J.*, vol. 11, no. 8, pp. 13108–13124, Apr. 2024.

[29] G. Chen, Y. Zhang, Z. Zhao, and F. Song, "QFA2SR:query-free adversarial transfer attacks to speaker recognition systems," in *Proc. 32nd USENIX Secur. Symp.*, 2023, pp. 2437–2454.

[30] M. AI, "Seamless communication," 2023. Accessed: 21 May, 2024. [Online]. Available: https://ai.meta.com/blog/seamless-communication/

[31] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 5998–6008.

[32] J. Wu et al., "On decoder-only architecture for speech-to-text and large language model integration," in *Proc. 2023 IEEE Autom. Speech Recognit. Understanding Workshop*, 2023, pp. 1–8.

[33] A. Bapna et al., "mSLAM: Massively multilingual joint pre-training for speech and text," 2022, *arXiv:2202.01374*.

[34] J. Wu, L. Liu, W. Bi, and D.-Y. Yeung, "Rethinking targeted adversarial attacks for neural machine translation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 12191–12195.

[35] S. Sadrizadeh, L. Dolamic, and P. Frossard, "Transfool: An adversarial attack against neural machine translation models," *Trans. Mach. Learn. Res.*, 2023.

[36] X. Zhang, J. Zhang, Z. Chen, and K. He, "Crafting adversarial examples for neural machine translation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 1967–1977.

[37] J. Ebrahimi, D. Lowd, and D. Dou, "On adversarial examples for character-level neural machine translation," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 653–663.

[38] C. Xu, J. Wang, Y. Tang, F. Guzmán, B. I. Rubinstein, and T. Cohn, "A targeted attack on black-box neural machine translation with parallel data poisoning," in *Proc. Web Conf.*, 2021, pp. 3638–3650.

[39] S. Sadrizadeh, A. D. Aghdam, L. Dolamic, and P. Frossard, "Targeted adversarial attacks against neural machine translation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[40] S. Poppi et al., "Towards understanding the fragility of multilingual llms against fine-tuning attacks," in *Proc. Findings Assoc. Comput. Linguistics: NAACL 2025*, 2025, pp. 2358–2372.

[41] F. Kreuk, Y. Adi, M. Cisse, and J. Keshet, "Fooling end-to-end speaker verification with adversarial examples," in *Proc. 2018 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 1962–1966.

[42] Z. Li, C. Shi, Y. Xie, J. Liu, B. Yuan, and Y. Chen, "Practical adversarial attacks against speaker recognition systems," in *Proc. 21st Int. Workshop Mobile Comput. Syst. Appl.*, 2020, pp. 9–14.

[43] Y. Xie, C. Shi, Z. Li, J. Liu, Y. Chen, and B. Yuan, "Real-time, universal, and robust adversarial attacks against speaker recognition systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 1738–1742.

[44] C.-X. Zuo, Z.-J. Jia, and W.-J. Li, "AdvTTS: Adversarial text-to-speech synthesis attack on speaker identification systems," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2024, pp. 4840–4844.

[45] W. Huang, J. Ren, H. Cao, H. Chen, H. Jiang, and Z. Fu, "Mitigating voice assistant eavesdropping via event source review on mobile devices," *IEEE Trans. Inf. Forensics Secur.*, vol. 20, pp. 9164–9179, 2025.

[46] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. 2017 IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[47] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. 2018 AAAI/ACM Conf. AI, Ethics, Soc.*, 2018, pp. 335–340.

[48] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Trans. Assoc. Comput. Linguistics*, vol. 6, pp. 587–604, 2018.
[49] L. Beinborn and R. Choenni, "Semantic drift in multilingual representations," *Comput. Linguistics*, vol. 46, no. 3, pp. 571–603, 2020.
[50] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction tuned LLM and latent diffusion model," 2023, *arXiv:2304.13731*.
[51] H. Liu et al., "Audioldm: Text-to-audio generation with latent diffusion models," in *Proc. Int. Conf. Mach. Learn. PMLR*, 2023, pp. 450–474.
[52] J. Melechovsky, Z. Guo, D. Ghosal, N. Majumder, D. Herremans, and S. Poria, "Mustango: Toward controllable text-to-music generation," in *Proc. 2024 Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (Volume 1: Long Papers)*, 2024, pp. 8293–8316.
[53] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10 695.
[54] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. 2015 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5206–5210.
[55] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. 16th Int. Conf. Digit. Signal Process.*, 2009, pp. 1–5.
[56] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," Dataset Univ. Edinburgh, 2019. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/3443
[57] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus," in *Proc. Interspeech 2021*, 2021, pp. 2756–2760.
[58] R. Ardila et al., "Common voice: A massively-multilingual speech corpus," in *Proc. Twelfth Lang. Resour. Eval. Conf.*, 2020, pp. 4218–4222.
[59] C. Wang et al., "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proc.59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process., Assoc. Comput. Linguistics*, 2021, pp. 993–1003.
[60] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. 2001 IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.
[61] E. Casanova et al., "Sc-glowtts: An efficient zero-shot multi-speaker text-to-speech model," in *Proc. Int. Speech Commun. Assoc.*, pp. 3645–3649, 2021.
[62] C. Liu, J. Zhang, T. Zhang, X. Yang, W. Zhang, and N. Yu, "Detecting voice cloning attacks via timbre watermarking," in *Proc. Netw. Distrib. System Secur. Symp.*, 2024.
[63] C. Jemine, "Resemblyzer," 2019. [Online]. Available: https://github.com/resemble-ai/Resemblyzer
[64] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., Assoc. Comput. Linguistics*, 2019, pp. 3982–3992.
[65] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet: A neural framework for mt evaluation," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process., Assoc. Comput. Linguistics*, 2020, pp. 2685–2702.
[66] Y. Liu, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[67] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding," in *Proc. 2018 EMNLP Workshop BlackboxNLP: Analyzing Interpreting Neural Netw. NLP*, 2018, pp. 353–355.

Chang Liu received the BS degree from the Hefei University of Technology, Hefei, China, in 2021. He is currently working toward the PhD degree with the University of Science and Technology of China under the guidance of Prof. Weiming Zhang. His research interests include AI security & privacy, watermarking, and audio processing.

Haolin Wu received the BE degree in information security in 2021 from Wuhan University, Hubei, China, where he is currently working toward the PhD degree with the School of Cyber Science and Engineering. His research focuses on the area of multimedia and machine learning security and privacy.

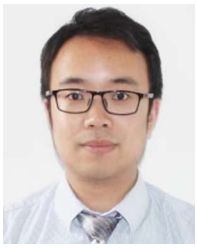
Xi Yang received the undergraduation and PhD degrees from the School of Cyber Science and Technology, University of Science and Technology of China. He is currently a postdoctoral fellow with the Hong Kong University of Science and Technology. His research interests include information hiding and AI security.

Kui Zhang received the PhD degree from the University of Science and Technology of China, Hefei, China, in 2024. He is currently a researcher with Huawei Noah's Ark Lab, Shanghai, China. His research interests include multi-modal LLM, visual reasoning, and trustworthy AI.

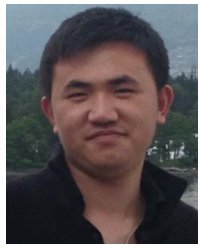
Cong Wu received the PhD degree from the School of Cyber Science and Engineering, Wuhan University in 2022. He is currently a research fellow with Cyber Security Lab, Nanyang Technological University, Singapore. His research interests include AI system security and web3 security. His leading research outcomes have appeared in USENIX Security, ACM CCS, *IEEE Transactions on Dependable and Secure Computing*, and *IEEE Transactions on Mobile Computing*.

Weiming Zhang (Member, IEEE) received the MS and PhD degrees from Zhengzhou Information Science and Technology Institute, China, in 2002 and 2005, respectively. He is currently a professor with the School of Information Science and Technology, University of Science and Technology of China. His research interests include information hiding and multimedia security.

Nenghai Yu received the PhD degree from USTC in 2004. He was a visiting scholar with the Institute of Production Technology, Faculty of Engineering, University of Tokyo, in 1999, and did cooperative research as a senior visiting scholar with the Department of Electrical Engineering, Columbia University, from April 2008 to October 2008. He is currently a full professor with the University of Science and Technology of China. He is also the director of Information Processing Center, USTC and the deputy director of Academic Committee, School of Information Science and Technology. His research interests include image processing and video analysis, multimedia communication, media content security, Internet information retrieval, data mining and content filtering, network communication and security.



Tianwei Zhang (Member, IEEE) received the bachelor's degree from Peking University, in 2011, and the PhD degree from Princeton University, in 2017. He is currently an associate professor with the College of Computing and Data Science, Nanyang Technological University. His research focuses on computer system security. He is particularly interested in security threats and defenses in machine learning systems, autonomous systems, computer architecture and distributed systems.



Qing Guo (Senior Member, IEEE) received the PhD degree from the School of Computer Science and Technology, Tianjin University, China. He was a research fellow and the Wallenberg-NTU Presidential postdoctoral fellow with Nanyang Technological University, Singapore, from 2019 to 2022. He is currently a senior research scientist and principal investigator with the Center for Frontier AI Research, A∗STAR, Singapore. He is also an adjunct assistant professor with the National University of Singapore. His research interests include computer vision, AI security, adversarial attacks, and robustness. He is the senior PC for AAAI and area chair for ICLR 2025.

Jie Zhang received the BS degree from the China University of Geosciences, Beijing, China, in 2017, and the PhD degree from the University of Science and Technology of China, in 2022. He is currently a research scientist with the Center for Frontier AI Research, Agency for Science, Technology and Research (A∗STAR), Singapore. His research interests include IP protection for AI, trustworthy generative AI, and AI regulation.