Adversarial Attacks Against Shared Knowledge Interpretation in Semantic Communications

Van-Tam Hoang, Van-Linh Nguyen, *Senior Member, IEEE*, Rong-Guey Chang, Po-Ching Lin, *Member, IEEE*, Ren-Hung Hwang, *Senior Member, IEEE*, Trung Q. Duong, *Fellow, IEEE*

***Abstract*—Semantic communications (SEMCOM) is a novel communication model that exploits neural networks or deep learning techniques to convey the semantics of the data and contextual reasoning, instead of transmitting full raw bits as in the conventional transmission models. SEMCOM is anticipated to significantly increase the effectiveness of cognitive communications beyond the Shannon theory limit, especially in multimedia services. The transmission efficiency will largely rely on the semantic encoding and decoding process with knowledge storage references at the receiver and the transmitter. However, these processes are highly susceptible to adversarial attacks, given the nature of shared background knowledge without encryption and the vulnerabilities of neural network models. This paper presents two novel targeted and non-targeted adversarial attacks against SEMCOM, e.g., channel inversion attack and naive attack. The attacks are designed to cause maximum disruption to the signals during decoding, aiming to alter the semantic interpretation of recognition models at the receiver. The experimental results indicate that attacks can significantly degrade the perceptual evaluation of speech quality and increase data errors, with semantic decoding performance suffering reductions of up to 2.9 times and 2.3 times, respectively. This degradation can cause misrepresentation of semantic contents. Besides, targeted attacks have a greater impact on speech semantic quality in complex communication circumstances compared to non-targeted attacks. We also suggest two potential defense methods against these physical layer attacks. Accordingly, enhancing adversarial training and removing residual values in the loss function are straightforward solutions to improve the resilience of SEMCOM-based systems.**

***Index Terms*—Wireless channel, Adversarial attack, Semantic communications, AI-based speech control.**

I. Introduction

Semantic communications is one of the emerging technologies for addressing extremely high bandwidth demands from ultra high-definition (UHD) or 4K/8K resolution video services and holographic content providers in the sixth-generation (6G) networks [1]–[4]. To increase the data rate, the research community is leaning on three options: (1) exploring higher wireless frequencies, such as Terahertz; (2) developing advanced modulation/physical coding techniques to compress data and use existing communication channels; (3) exploring artificial intelligence (AI)'s power to enable new cognitive communications where context 'updates' or new knowledge on the situation are transmitted only. In the first and second transmission models, the data bits are sent across wired or wireless channels for transfer to the recipients. The maximum rate of a wireless channel is frequently influenced by physical channel parameters such as channel bandwidth, carrier frequency, and fading/noise. The rate is bounded by the Shannon limit [3], i.e., there is a restriction on the quantity of data sent by a signal with a certain amount of power.

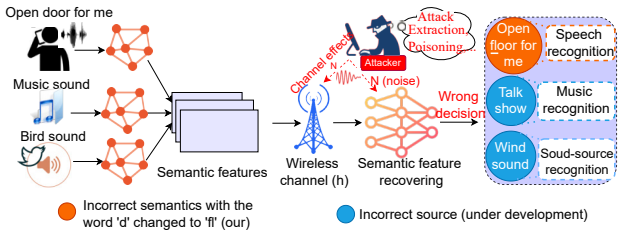

Fig. 1. Illustration of an adversarial attack by creating perturbations on semantic communication systems to affect the semantic quality. After being targeted, voice recognition/classification systems in AI-related applications will be misrecognized/misclassified semantically.

The third transmission model represents a novel approach that exploits context extraction through natural language processing techniques and generative AI models to convey the meaning of messages from the transmitter to the receiver. For example, instead of transmitting bit sequences representing a full UHD video and voice of "the kid in the kitchen", semantic communications only extracts the *personalized text* "the kid in the kitchen" and transmit this necessary information to the receiver. The full video can be reconstructed by generative AI models (e.g., OpenAI Sora [5], CV-VAE [6]) based on the sharing of stored knowledge between the transmitter and the reviewer. The irrelevant information, such as the home background (which is already known by the transmitter and the receiver through the knowledge database), will be skipped to decrease the amount of transferred data while maintaining performance. Therefore, semantic communications will considerably decrease the amount of raw data to pour into network channels, e.g., speech transmission [3], [4], image transmission [7], [8], and video transmission [4].

However, the major challenges of semantic communica-

This work was supported in part by the National Science and Technology Council (NSTC) of Taiwan under Grants 112-2221-E-194-017-MY3 and in part by the Advanced Institute of Manufacturing with High-tech Innovations from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan. The work of T. Q. Duong was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109.

Van-Tam Hoang, Van-Linh Nguyen, Rong-Guey Chang, and Po-Ching Lin are with National Chung Cheng University (CCU), Taiwan, and the Advanced Institute of Manufacturing with High-tech Innovations, CCU, Taiwan.

Ren-Hung Hwang is with the College of Artificial Intelligence, National Yang-Ming Chiao Tung University (NYCU), Taiwan.

Trung Q. Duong is with Memorial University, Canada.

Corresponding author: Van-Linh Nguyen (email: nvlinh@cs.ccu.edu.tw).

tions are the nature of shared background knowledge without encryption (AI can understand) and the vulnerabilities of neural network-based semantic encoding models to adversarial attacks [9]. Fig. 1 illustrates an example of adversarial attacks on semantic communication systems to degrade the semantic quality. As illustrated in Fig. 1, the ultimate goal of the communication process is to transmit semantics from the transmitter to the receiver through multiple path channels. The attackers aim to deploy adversarial attack methods to intercept information or cause misrecognition/misclassification from semantic decoding by exploring the vulnerabilities in the SEMCOM encoding process. For semantics, the encoding/decoding process is vulnerable to adversarial attacks in several ways, e.g., knowledge store poisoning and channel-aware adversarial attacks [10], [11]. Adversaries can alter the input data or insert hidden triggers under the type of bits with text, signals with speech, and symbols of orthogonal frequency division multiplexing (OFDM) to the deep neural networks (DNNs)-based semantic communications. These outputs will be reconstructed at the receiver, which includes both original input and perturbation. Once the adversarial bits, signals, and symbols of OFDM are reconstructed with clean input, the final result is changed to misinterpret the semantic content. Accordingly, voice recognition systems misclassify human commands as a result of manipulated and unrecognizable perturbations in input data. Many existing studies in semantic communications rely on DNNs and advanced generative AI models, such as GPT's and LLM's models [12], to extract semantic features and transmit them to the receiver. However, DNNs have been a well-known target for many studies on adversarial attacks [13]. Semantic communications are also in the early stages of development. Adversarial attacks in semantic communications have not yet been well-explored.

This study presents novel physical-layer adversarial attacks against semantic communications through wireless channels. The proposed method selectively transmits semantic features that are important for the intended transmission tasks and penetrates the efficiency of several perturbation patterns in various transmission contexts, e.g., line-of-sight (LoS) and non-line-of-sight (NLoS). The perturbation patterns can cause misinterpretation of semantic-based automated speech recognition (ASR) in smart home applications or smart speakers. The main contributions are presented briefly as follows.

1) The work is the first attempt to design the real end-to-end semantic interpretation and extraction of a semantic communication-based speech service in the wireless channel. The system's vulnerabilities in the semantic interpretation process are also analyzed for designing adversarial attacks. Implementation of the attacks and a sample defense can provide a valuable starting point for future studies.
2) This is the first attempt to propose two efficient adversarial attacks, the targeted and non-targeted attack, that aimed at degrading the quality of semantic communications in the semantic interpretation process. The semantic quality in decoding is significantly reduced up to 2.9 times. Specifically, the signal-to-distortion scores of the targeted attack on the AWGN, Rayleigh, and Rician channels were reduced to 1.35, 2.68, and 3.02 times, respectively, compared to non-targeted attacks.
3) This study also provides a comprehensive evaluation regarding the semantic interpretation that changes the semantic quality before and after the attack on the proposed system. Based on this analysis, we briefly describe the defense mechanisms against adversarial speech attacks. Generally, the study aims to evaluate and safeguard the integrity and meaningfulness of communication despite potential attempts to disrupt or manipulate the transmitted information.

The remainder of the study is organized as follows. Section II discusses the related works. Section III describes the proposed SEMCOM system model. Section IV presents the proposed physical-layer adversarial attack process. Sections V and VI present semantic-targeted and non-targeted adversarial attacks on wireless channels, respectively. Section VII evaluates the attack performance for the proposed system. Section VIII concludes the paper.

II. Related works

This section introduces related works on state-of-the-art SEMCOM techniques and preliminary studies of adversarial attacks in SEMCOM. The channel factors influencing adversarial perturbations are also discussed. Table I summarizes several typical studies on semantic communications, key features, major contributions, influence factors, as well as semantic quality evaluation compared to this study's research position.

A. End-to-End semantic communication systems and preliminary studies of adversarial attacks

There are several studies on semantic communications and their security vulnerabilities in the literature. For example, the study in [14] presented the semantic error minimization process by merging semantic inference and the physical layer on both the transmitter and receiver. The authors of [15], [16] developed a framework employing an edge server to categorize pictures and decrease the required transmission bandwidth. The authors referred to these properties as semantics after the system has extracted data from the input data. However, the attacks on the systems to alter their semantics in wireless channels have yet to be realized. A study in [17] proposed a semantic attack on the physical layer. The authors focuses on evaluating the semantic error rate in picture data poisoning and provide little information on the contexts of wireless channel sensitivity for semantic quality.

The study in [24] experimented with a speech semantic system for indoor THz-wireless communications. However, the authors did not address adversarial attacks or the influence factors of wireless channel sensitivity on semantic quality. Other studies in [3] and [14] examined the efficiency of semantic communication systems for text and speech data. However, the effect of adversarial attacks on each wireless channel that impacts semantic quality has yet to be examined. The authors have solely looked at the semantic error rate. In the research [18], the authors proposed a robust system

TABLE I
COMPARISON OF RELATED WORKS ON SEMANTIC QUALITY AND CHANNEL EFFECTS UNDER ATTACKS

Study	Dataset	Attack	Defense	Attack performance	Channel effect	Semantic quality	Measurement metrics	Major contribution	Limitation
[1]	Text	✗	✗	✓	✗	✓	BLEU	Evaluate semantic quality on Internet of Things (IoT) devices using channel state information.	Ignore adversarial attacks and character noise.
[3]	Speech	✗	✗	✗	✗	✓	SDR, PESQ	Evaluate the recoverability of semantic features via SDR and PESQ scores.	Not considering perturbation under adversarial attacks.
[14]	Text	✗	✗	✗	✗	✓	BLEU	Minimize the semantic errors by recovering the meaning of sentences under BLEU score.	Ignore channel effects and character noise.
[17]	Image	✓	✓	✗	✗	✓	PSNR, SSIM	Evaluate the physical-layer adversarial robustness under GAN attacks in SEMCOM systems.	High computation complexity while training adversarial.
[18]	Text	✗	✗	✓	✗	✗	BLEU Score	Evaluate the robustness of the semantic communication systems in terms text.	Not addressing the text corruption due to channel effects.
[19]	Radio	✓	✓	✗	✓	✗	AC	Provide significant insights into the susceptibility of modulation classifiers to adversarial attacks.	Not cover all possible real-world channel conditions.
[20]	Signals	✓	✗	✗	✗	✗	AC	Evaluate the performance of adversarial attacks to detect channel effects against modulation classifiers via a deep learning model.	Not considering the semantic quality on each channel.
[21]	Radio	✓	✗	✓	✗	✗	AC	Evaluate the efficiency of the white-box and black-box adversarial attacks against the modulation classifiers for radio signals.	Not considering the semantic quality as well as the channels in the system.
[22]	Video	✓	✗	✓	✗	✗	FPR, PCK, AC, TPR	Consider forgery attack detectability on supervised video streams in the backdoor camera using the channel state information.	The semantic quality was not considered or discussed.
[23]	Signal	✗	✗	✗	✗	✓	PESQ, SDR	Evaluate the efficiency of the semantics when transmitting signals on a wireless channel.	Not considering adversarial attacks affecting semantic quality.
Our paper	Speech	✓	✓	✓	✓	✓	SDR, PESQ, DRE	- Attack the mean squared error (MSE) loss functic to change the semantic quality in SEMCOM systen - Discuss corresponding defense strategy to counter adversarial attacks.	Target specific multimedia service and open source AI platforms.

AC: Accuracy; **BLEU**: Bilingual evaluation understudy; **PSNR**: Peak signal-to-noise ratio; **SSIM**: Structural similarity index measure; **BER**: Bit-error rate; **FPR**: False positive rate; **TPR**: True positive rate; **PCK**: Percentage of correct keypoint; **DRE**: Data Rate Error; ✓ is measured, ✗ is opposite.

for transmitting semantic voice data. However, the study did not address the effect of artificial noise from intentional attacks. The study in [25] considered a comprehensive error performance comparison among wireless channels. However, this comparison focused on static channels instead of the input of dynamic semantic communications. In short, none of the aforementioned studies are specified for adversarial attacks against semantic communications with personalized contexts, let alone exploit various characteristics of the wireless channels to directly impact the semantic quality.

B. Semantic data with the influence of channel sensitivity on adversarial attacks

In SEMCOM, the attacker can listen to penetrate the channel characteristics and related data, such as sampling rate, input data type, and corresponding output. For example, many studies on adversarial machine learning [9], [26] indicate that channel sensitivity knowledge can be crucial to the success rate of adversarial samples. The studies in [27], [28] also highlight DNN-based techniques to learn the conveyed ability in the channel models and generate proper channel noise. The methods' successful attack rates indicate credible performance. However, there is a lack of discussion on channel sensitivity on the impact of semantic quality in media services. The authors in [29] proposed a new spectrum poisoning attack, where the attacker can falsify a transmitter's spectrum sensing data over the air by transmitting adversarial noises during the spectrum sensing period of the transmitter. Based on the baseline, many studies expand the research to specific objects such as spectrum sensing [30], [31] and IoT data aggregation [32]. Besides, there are several studies on adversarial trojan attacks in wireless communications [9], waveform [33], and channel [34]. However, these works focused on a modulation classifier and did not discuss channel sensitivity and transmission defects in an end-to-end semantic communication system.

Adversarial machine learning and attack techniques in the fifth-generation (5G) networks have been studied in [35], [36] to deploy spoofing attacks based on DNNs to fool signals or disrupt channel authentication systems. Several attack techniques in emerging communication techniques, such as 6G and THz have also been studied and show promising results [37]. However, evaluating semantic quality in various channel models (e.g., AWGN channel) in semantic communication systems has still not been considered, especially in intelligent speech recognition services or 6G-related semantic applications.

III. SEMANTIC COMMUNICATION SYSTEM MODEL AND ATTACK MODEL

Fig. 2 depicts a SEMCOM model structure that contains three key parts: a transmitter, a wireless channel model, and a receiver. The parts can be simulated by the convolutional neural network (CNN). This work also uses this approach to model the weight of speech transmission. In the channel model, the semantic knowledge base is stored as weights in CNN models (e.g., ResNet) after training with different audio/voice datasets. The details of the three parts of the SEMCOM model are presented in the following subsections.

A. Transmitter modeling

As the communication diagram of semantic-based speech transmission systems illustrated in Fig. 2, the transmitter includes two separate components, a *semantic encoder's* CNN and a *channel encoder's* CNN. Assume that the learning parameters of the semantic's encoder and channel's encoder are δ and α, respectively. The semantic transmission process

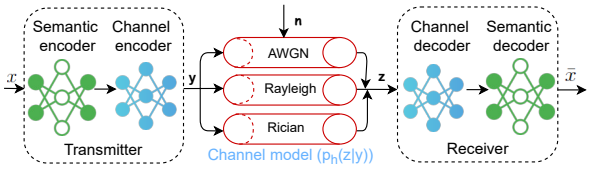

Fig. 2. The model structure of a SEMCOM system with three channel models (different noise/transmission scenarios).

of the transmitter part is as follows. First, the input of the transmitter is a clean speech sample framing sequence, $x = [x_1, x_2, \ldots, x_M]$, with M samples, where x_m is m^{th} element in x. Second, these input speech samples are mapped and encoded to symbol sequence y which becomes the input of the wireless channel and is given by

$$y = C_\alpha(S_\delta(x)), \tag{1}$$

where $C_\alpha(.)$ and $S_\delta(.)$ are the *channel encoder* and *semantic encoder* with respect to (w.r.t.) parameters α and δ, respectively. We use $\theta^t = (\alpha, \delta)$ to represent the CNN's parameter of the transmitter, which will become the input of the wireless channel. Before being transmitted to the wireless channels, the altered data must be normalized to ensure that the transmitted power remains at a constant value. The normalization formula is computed by $P_{max}\|y\|^2 = 1$.

B. Semantic communication channel modeling

Fig. 2 illustrates a single semantic communication channel between the transmitter and the receiver. Unlike the conventional communication model with a full TCP/IP protocol stack, this study assumes that SEMCOM uses built-in channel encoders and channel decoders *without* modulation and physical layer encryption, which is common in pure AI-based wireless communications [38]. Therefore, adversarial attacks against speech recognition during the encoding and decoding process will not be validated by integrity checks as in conventional communication models. This study explores three cases of communication with three corresponding channels: AWGN channel, Rayleigh channel, and Rician channel. AWGN channel indicates the impairment to communication is a linear addition of wideband or white noise with a constant spectral density. The Rayleigh channel denotes the magnitude of a signal that has passed through such a communication channel will vary randomly, or fade, which is the radial component of the sum of two uncorrelated Gaussian random variables. Rician channel means there is a LoS propagation dominating multipath components. In short, the channel model, denoted by $p_h(z|y)$, will take y as the input and produce the output as received signal z. As presented in Fig. 2, we may model the signal transmission process from transmitter to receiver through the channels, which can be generally described by

$$z = h * y + g, \tag{2}$$

where h is the linear channel coefficient. $g \sim \mathcal{CN}(0, \sigma^2 I)$ is the Gaussian noise, σ^2 is the perturbation variations for each channel and I is the identity matrix.

C. Receiver modeling

Similar to the transmitter model, the receiver also consists of two cascaded parts, including the *channel decoder* and the *semantic decoder* as illustrated in Fig. 2. To fit the computation process, we assume that η and ω are the CNN parameters of the semantic and channel decoders, respectively. With the communication scenario from the output z, the received output signal $\overline{x}$ will be expressed by

$$\overline{x} = S_\eta(C_\omega(z)), \tag{3}$$

where $S_\eta(.)$ is the semantic decoder, and $C_\omega(.)$ is the channel decoder. Its received CNN parameter set is $\theta^r = (\eta, \omega)$. The final result of the communication system is that the speech output signals will be decoded at the *channel decoder* and reconstructed at the *semantic decoder* to be as close to the original input format as possible. However, knowing "whether the signals are manipulated to add new perturbations before decoding" will be difficult. The attacker can inject adversarial samples and the system will decode and reconstruct both original signals and adversarial perturbated signals.

D. Threat Model

Adversarial attacks on semantic communication systems are critical and feasible in several contexts. For example, the authors in [39] created a novel approach for crafting physical layer black-box adversarial attacks for SEMCOM systems. As a result, the method can sharply decrease the classification accuracy. The significant loss of communication efficiency by only transmitting incorrect data shows potential risks for applications like automatic driving, digital twins, and smart health. Since the SEMCOM is based on neural networks, the communication paradigm will focus on how the transmitted symbols convey the desired meaning [40], instead of accurate bits. In this work, we tamper the semantic decoder at the device with adversarial signals. This attack threat is possible since data is not encrypted before the AI-based devices perform encoding. For example, when a human speaks to a smart speaker device, the voice is recorded and then semantically encoded on the device. Semantic encoding at the smart device can be misled by intentional noise and adversarial signals. The neural network-based channel estimation for SEMCOM can be also the target of adversarial signals.

IV. Proposed Physical-Layer Adversarial Attacks

This section explains the steps to build physical-layer adversarial attacks on wireless channel semantics. This work assumes the generated perturbations are synchronized with the transmitter's signals, i.e., perturbation ϕ_k and input signal x have $2n$ dimension vectors. As a result, each element of ϕ_k must be added to the corresponding element of x to obtain $x * \mathcal{H} + n + \phi_k$.

A. Proposed semantic architecture and workflow

Fig. 3 presents the proposed architecture and system model under adversarial attacks. The input data are voice sample sequences, $\mathbf{X} = \Im^{b \times M}$, where $\mathbf{X}$ consists of the voice data

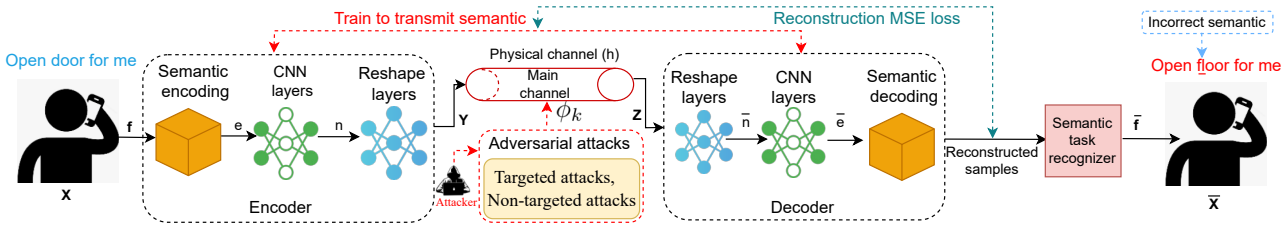

Fig. 3. The overview of our proposed attacks on SEMCOM, where the meaning of input signals is altered due to adversarial attacks.

TABLE II
THE DETAILED MODEL STRUCTURE FOR THE PROPOSED SYSTEM.

Component	Layer	Filters	Action
Transmitter	6×SE-ResNet model	384	Relu layer
	CNN layer	8	Conv. layer
Wireless channel	AWGN, Rayleigh, and Rician channel	None	None
Receiver	CNN layers	8	Conv. layer
	6×SE-ResNet models	384	Relu layer
	Last Layer (CNN)	1	Reshape layer

set x and b is the batch size of the input. Note that voice samples $\mathbf{f}$ are framed with a fixed length for encoding. While data transfers pass the semantic encoder, these voice signals will be encoded. Semantic and outputs will be $\mathbf{e} \in \Im^{b\times k\times l\times d}$, where k and l are the number and length for each voice sample frame, respectively. d is the dimension of each frame. The CNN layer is channel encoder and encode $\mathbf{e}$ to $\mathbf{n} \in \Im^{b\times k\times 2N}$, and $\mathbf{n}$ will be reshaped into symbol sequences and forward to $\mathbf{Y} = \Im^{b\times kN\times 2}$, and then $\mathbf{Y}$ is the input of wireless channels. These channels will receive the reshaped symbol sequences from $\mathbf{Y}$ and its output becomes the input of $\mathbf{Z}$ at the receiver. In this study $\mathbf{Z}$ is computed by

$$\mathbf{Z} = \mathcal{H} * \mathbf{Y} + \mathcal{N}, \tag{4}$$

where $\mathcal{H}$ is the channel fading coefficient vectors of h and $\mathcal{N}$ refers to the Gaussian noise vectors of g. The receiver's decoder decodes the output signals $\mathbf{Z}$ with similar sizes to the input, denoted by $\bar{\mathbf{n}}$, $\bar{\mathbf{e}}$, and $\bar{\mathbf{f}}$, and finally reconstructs $\bar{\mathbf{f}}$ into $\overline{\mathbf{X}}$ via the inverse operation of framing. The semantic task recognizer is an application/system that is used to recognize the reconstructed signals at the receiver, e.g., the input is "Open door for me" while the output is "Open *f*loor for me" (as in Fig. 3). The training model structure to transmit signals from the transmitter and receiver is outlined in Table II. The training and noise injection process in this work is briefly described in the Algorithm 1.

Algorithm 1: Noise injection and attack training

Input: Speech data set X, noise P_n from attack method below.
Result: Trained networks $C_\alpha(.)$, $S_\delta(.)$, $S_\eta(.)$, $C_\omega(.)$
Convert speech (X) to framing f to fit with train.
while *Stop criterion does not meet* **do**
  $C_\alpha(\mathbf{f}) \rightarrow \mathbf{e}$
  $S_\delta(\mathbf{e}) \rightarrow \mathbf{Y}$
  Transmit Y to physical channels + noise (ϕ_k) and receiver $\mathbf{Z} = Y + \phi_k + \mathcal{N}$.
  $S_\eta(\mathbf{Z}) \rightarrow \bar{\mathbf{e}}$
  $C_\omega(\bar{\mathbf{e}}) \rightarrow \bar{\mathbf{f}}$
  Convert $\bar{\mathbf{f}}$ to framing $\overline{\mathbf{X}}$
  Compute parameters $\boldsymbol{\theta} = (\sigma^t, \sigma^r)$
  Recompute trainable parameters and add noise
  Update parameters and noise (*noise-params, X)
  $L_k^t(\theta^t) = \frac{1}{K}\sum_{k=1}^{K}(x_k - \overline{x}_k)^2 + P_n$
  $L_k^r(\theta^r) = \frac{1}{K}\sum_{k=1}^{K}(x_k - \overline{x}_k)^2 + P_n$
  $k \rightarrow k+1$
end

B. The stages of adversarial attacks and problem formulation

The objective of this work is to generate perturbation signals and merge them with semantic signals in wireless channels. The mixed signals can mislead the recognition engine at the receiver and lead to wrong conclusions. The processes are summarized in the following stages.

(1) In the first stage, the transmitter sends speech signal frames that contain the semantics of the speech to a wireless channel, where each frame contains one semantic class, such as "Open door for me" in a smart home as shown in Fig. 3.

(2) The data is then enhanced with additive white Gaussian noise and adversarial perturbations, i.e., perturbation ϕ_k, where the first can be called natural perturbation, whereas second perturbations attempt to assault the semantics, e.g., the semantic meaning of the word "Open door for me". This injection is feasible because the speech signal frame is always mapped or transmitted under bit-level weight matrices on wireless channels. The perturbations are mapped to channels via weight matrices, too. In this work, the adversarial perturbation ϕ_k is assumed to be synced with the transmitter's signal and overlaid on the transmitted signal x at the receiver. The attacker then makes replay attacks on the channels to change the semantic quality at the receiver. To make it more practical, the perturbations produced by the generator are input-agnostic.

(3) The receiver either receives signals directly from the recognizer or reconstructs all received signals before sending them to the recognizer, which outputs predictions based on semantic information. In this study, we consider both targeted and non-targeted attacks (detailed in the next sections), where the signals sent to the channel contain both random and targeted signals. The reconstructed signals, with added pertur-

bations, will fool the system into making incorrect decisions. As simulated in Fig. 3, the phrase "Open door for me" can be misrecognized as "Open **fl**oor for me" in a smart home under adversarial attacks. This demonstrates that these signals have been injected with a fixed perturbation signal that is hard for humans to detect. The perturbation generation process mainly relies on the adversarial perturbations ϕ_k for the signal, which is computed by

$$\begin{aligned} &\arg\min_{\phi_k} \|\phi_k\|_2^2 \\ &\text{subject to} \quad Z(x_k, \phi_k) \neq Z(x_k), \forall x_k \\ &\|\phi_k\|_2^2 \leq \epsilon \end{aligned} \tag{5}$$

where x_k is the input signal k^{th} and ϵ is a hyper parameter for the perturbation's upper-limitation, $\|\phi_k\|_2^2$ is the squared L_2-norm of ϕ_k. However, the adversarial perturbation ϕ_k is restricted by the power budget computed from ϵ. It is important to note that the optimal solution is not always reached when $\|\phi_k\|_2^2 = \epsilon$ due to the DNN's complicated decision boundaries, which is determined by the power and phase of the disturbance. Therefore, solving the optimization problem in Equation (5) is challenging, and its complexity is increasing today due to the nonlinearity and convexity of the DNNs. There are various strategies developed, particularly in the computer vision field, to estimate these adversarial examples. For example, the Fast Gradient Method (FGM) [41] is an efficient way to create adversarial attacks by linearizing the loss function in the neighborhood of input x of a DNN classifier. A key challenge in Equation (5) is its focus on attacking content-oriented wireless networks while neglecting semantic attacks, where the receiver relies on semantics for accurate inference or action. This work introduces a novel mechanism to generate perturbations and degrade semantic quality in semantic-oriented wireless networks, detailed in the following subsection.

C. The new perturbation generator for semantic networks

In the SEMCOM-based speech system, the human voice is the source of the signals. Due to the different tone characteristics of each human speaker, the attacker can easily capture their speech signals and replay them. This work assumes that the replayed signals are synchronized with the transmitter's original signals. The attacker can mislead the speech recognition models by replaying the gathered signals from the victim with perturbation patterns via an AI-based virtual assistant, voice synthesizer, or when the victim is absent (smart home). To conceal the attack factors, the attacker can embed the noise generator in the devices connected to the system, such as smartphones, speech recognition devices, and compromised applications. In our system, semantics in the communication system will be sent to the receiver as illustrated in Equation (4). After attack, Equation (4) is rewritten as

$$Z(\phi_k) = \mathcal{H} * \mathbf{Y} + \phi_k + \mathcal{N}, \tag{6}$$

where ϕ_k is the maximum perturbation between the original sample x and the adversarial example according to attack strategies. The main goal of the whole system focuses on attacking semantics on channel models and evaluating the semantic quality under attacks at the receiver. The accuracy of the decoding process depends on the knowledge base of the channel/semantic decoder, as well as the cleanliness of the input voice data. The update process always depends on the loss function during training. This also means that our speech signal characteristics between x and $\overline{x}$ in Equation (4) are also evaluated via mean-squared error (MSE) loss function L_{mse} [3]. In detail, the loss function of k^{th} element in the proposed semantic system is given by

$$\begin{aligned} L_{mse}(\theta_k) &= \frac{1}{K}\sum_{k=1}^{K}(x_k - \overline{x}_k)^2 = \sum_{k=1}^{K} P(x_k, \theta_k) \\ \text{with } P &= \frac{1}{K}(x_k - \overline{x}_k)^2, \end{aligned} \tag{7}$$

where x_k and $\overline{x}_k$ represent the input and the output signal, respectively. K is the length of vectors x and $\overline{x}$. Assume that the NN models of the entire transceiver are differentiable in terms of the appropriate parameters and can be tuned using gradient descent based on Equation (8). Note that the semantic and channel encoder/decoders are designed together. Also, with given previous parameters of CSI, the system can adjust both NN parameter sets σ^t, σ^r at the same time. Therefore, we call the NN parameter set of the whole system, $\theta = (\sigma^t, \sigma^r)$. From study [3], the NN parameter θ in the loss function of the system will update iteratively, which is expressed by

$$\theta^{k+1} = \theta^k - \rho \bigtriangledown_{\theta_k} L_{mse}(\theta_k), \tag{8}$$

where $\rho > 0$ is a learning rate and $\bigtriangledown$ indicates the differential operator. As shown in Fig. 3, the reconstructed signals are based on Equation (8). Therefore, we attack the channel by adding perturbation ϕ_k to Equation (6). To find perturbation ϕ_k, this problem can be transferred to a constrained optimization problem as in study [42]. Therefore ϕ_k is satisfied by

$$\begin{aligned} &\arg\max_{\phi_k} \|\phi_k\|_2^2 \\ &\text{subject to} \quad L_{\text{MSE}}(x_k, \phi_k, \theta_k) \neq L_{\text{MSE}}(x_k, \theta_k), \end{aligned} \tag{9}$$

where $L_{\text{MSE}}(x_k, \theta_k)$ is the original loss function, $L_{\text{MSE}}(x_k, \phi_k, \theta_k)$ is the loss function after adding perturbation ϕ_k. By combining Equations (7) and (9) for perturbation element k^{th}, it is expressed by

$$\sum_{k=1}^{K} P(x_k, \phi_k, \theta_k) \neq \sum_{k=1}^{K} P(x_k, \theta_k) \tag{10}$$

Because the recognizer is sensitive both the direction and the power of perturbation, the squared error criterion will penalize the candidates of ϕ_k that have more power with the direction of the original example x. We set $\phi_k = \gamma \times x_k$ to search for all magnitudes of the ϕ_k, where γ is a path loss coefficient that can be optimized by line search. According study [19], the distance between the original sample x_k and the adversarial example ϕ_k is given by

$$\sum_{k=1}^{K}(\phi_k - \gamma x_k) \tag{11}$$

The main target is to maximize the perturbation on the loss function, assuming that the intruder has limited signal resources. In other words, the attacker only has ϵ chances to launch attacks each time. To increase the attack performance as much as possible, attackers seek to maximize the cost function. Therefore, Equation (11) is expressed by

$$\begin{aligned} &\max_{\phi_k} \sum_{k=1}^{K} \|\phi_k - \gamma x_k\|_2^2 \\ &\text{subject to} \quad \sum_{k=1}^{K} \|\phi_k\|_2^2 \leq \epsilon \end{aligned} \tag{12}$$

To maximum Equation (12), this study uses the Lagrangian method [43] to optimize the convex problem. The Lagrangian for (12) is given by

$$L_{MSE} = \sum_{k=1}^{K} \|(\phi_k - \gamma x_k)\|_2^2 + \lambda \left(\sum_{k=1}^{K} \|\phi_k\|_2^2 - \epsilon \right), \tag{13}$$

where $\lambda \geq 0$, and the Karush–Kuhn–Tucker (KKT) conditions to get maximum value in (12) are expressed by

$$\phi^*(\phi_k - \gamma x_k) + \lambda \phi_k = 0, \text{with } k = 1, ..., K \tag{14}$$

From Equation (14), we can get the maximized perturbation of L_{MSE} for Equation (6), which is computed by

$$\phi_k = \frac{\phi^* \gamma x_k}{\phi^* + \lambda}, \tag{15}$$

where ϕ^* is the conjugate of perturbations, λ is determined by the adversary's power constraints. Designing perturbations in this manner ensures that the received disturbance matches the input signals while adhering to the adversary's power limits.

D. Complexity analysis

To evaluate the computation complexity of the proposed system, we quantify the complexity of the CNN-based model in terms of the number of floating-point operations (FLOPs) performed by the convolutional kernels and evaluate Algorithm 1 in terms of its time and space complexities. First, to compute the FLOPs of the CNN-based model, which is used to transmit semantics in wireless channels, we calculate FLOPs for a single 2D CNN module by [44]

$$C_{2DCNN} = 2 \times w \times h \times (c_{in} \times k^2 + 1) \times c_{out}, \tag{16}$$

where w is the width and h is the height of CNN's input feature map. k represents the kernel size. c_{in} is the number of the input layers and c_{out} is the number of the output layers[1] of feature maps. Based on Equation (16), we summarize the FLOPs of the CNN-based model in this study and the traditional model, as shown in Table III. We found that the CNN-based systems require a higher computational cost than conventional techniques. This is due to the complexity of neural network training and the incorporation of the feature encoder/decoder process during model training. With the traditional system, the feature encoder takes speech samples as inputs, and its output is supplied directly to the decoder. The received signals are converted into semantic information.

[1]Here, the layers refers to the parameter of CNN

TABLE III
COMPARISON OF FLOPS IN TRADITIONAL, CNN-BASED SYSTEM

Component	Traditional system (1)	CNN-based system (2)	SE-ResNet (our)	Change FLOPs (1) ——— (2)
Transmitter	2.76×10^9	4.45×10^9	4.65×10^9	$1.69 \times 10^9 \Uparrow -1.89 \times 10^9 \Uparrow$
Wireless channel	None	None	None	None
Receiver	2.82×10^9	4.50×10^9	4.72×10^9	$1.68 \times 10^9 \Uparrow -1.9 \times 10^9 \Uparrow$

Note: The model structure of the traditional system follows that of [3], while the other one is detailed in Table II.

Based on the input and output information, the MSE loss is computed at the receiver, and the trainable parameters of both the feature encoder and the feature decoder are updated simultaneously. Therefore, their computational complexity is lower than that of the CNN-based system, such as ours. Second, to evaluate the time complexity of Algorithm 1, we evaluate the time complexity in each trained network. With L_α, L_δ, L_η, L_ω representing the number of layers, and n_α, n_δ, n_η, n_ω representing the number of neurons per layer of each trained network $C_\alpha(.)$, $S_\delta(.)$, $S_\eta(.)$, $C_\omega(.)$, respectively. N is the sample size and T is the number of iterations in the while loop. The time complexity in four trained networks is approximately $O(Time) = O(N \times T \times (L_\alpha n_\alpha^2 + L_\delta n_\delta^2 + L_\eta n_\eta^2 + L_\omega n_\omega^2))$. The assignments and algebraic operations have approximately complexity $O(N)$. Therefore, the time complexity of Algorithm 1 is $O(Time) + O(N)$ for training a single model. For space complexity, storing the parameters in Algorithm 1 mainly is the liner computation. Therefore, space complexity is approximately $O(N + |\theta|)$, where $|\theta|$ is the total number of trainable parameters in the networks.

V. SEMANTIC-TARGETED ADVERSARIAL ATTACKS IN WIRELESS NETWORKS AND CALIBRATIONS

This section mathematically evaluates the semantic quality of a system by analyzing channel effects on speech, a factor overlooked in previous studies. Other signals, like natural sounds, are treated as natural noise affecting semantic communication. Using the system from Section III, the process involves: (1) inputting a predefined dataset, (2) transmitting framed speech signals through three wireless channels (Fig. 3), and (3) attacking the signals as outlined in Algorithm 1. This study customizes targeted adversarial attack techniques—fast gradient sign method (FGSM) and projected gradient descent (PGD)—for semantic communications. Unlike their standard computer vision versions, these techniques use a new perturbation generator and are organized based on their impact on random channel effects, detailed in the following subsections.

A. Fast gradient sign method with generated perturbations

The FGSM is known as an adversarial attack technique [45] through the optimization of a neural network for the loss function $L(f(x), y)$, where $f(x)$ is a function of the neural networks, x and y are the original input and its real target, respectively. The adversarial signals x^* created from input x are computed by

$$x^* = x + \delta sign(\bigtriangledown_x L(f(x), y)), \tag{17}$$

where $\bigtriangledown_x L(.)$ is the derivative of the loss function L on x, $sign(.)$ is the sign operation, δ is the attacking intensity parameter, and x^* is the adversarial sample on x.

Algorithm 2: Channel inversion attack and bisection search in semantic communications

1 **Input:** Voice data set X, desired accuracy λ_{acc}, power constraint value P_{max}, C is the number of samples and recognition model (M). Channel information h of vector H.
Result: Adversarial perturbation of the input, N^{adv}.
2 Initialize: $\lambda_{acc} = 0$
3 **for** *c in range(C)* **do**
4 $\quad \lambda_{max} = P$
5 $\quad \lambda_{min} = 0$
6 $\quad \delta_{norm} = \frac{H.\bigtriangledown_x M(\theta, h_i, z)}{\|H.\bigtriangledown_x M(\theta, h_i, z)\|_2}$
7 $\quad$ **while** $\lambda_{max} - \lambda_{min} > \lambda_{acc}$ **do**
8 $\quad\quad \lambda_{avg} = (\lambda_{max} + \lambda_{min})/2$
9 $\quad\quad x_{adv} = channel - inver(x, \lambda_{avg}, \delta_{norm})$
10 $\quad\quad$ **if** $attack(x_{adv}) = true$ **then**
11 $\quad\quad\quad \lambda_{min} = \lambda_{avg}$
12 $\quad\quad$ **end**
13 $\quad\quad$ **else**
14 $\quad\quad\quad \lambda_{max} = \lambda_{avg}$
15 $\quad\quad$ **end**
16 $\quad$ **end**
17 $\quad \lambda[c] = \lambda_{max}$
18 **end**
19 $target = argmin(\lambda_{acc})$
20 $N^{adv} = -\sqrt{P_{max}} \times \delta_{norm}[target]$
21 **Return** N^{adv}

B. Channel inversion attack and bisection search

Unlike the FGSM attack, the channel inversion attack finds the best-targeted attack with the least amount of disruption by using a bisection search to find the scaling factor. This technique is to make sure that misclassification happens within the perturbation norm constraint, and then causes misrecognition. The attacker can optimize the targeted adversarial signals N^{adv} to mitigate semantic quality, which is obtained by using Algorithm 2. First, the algorithm creates the normal perturbation δ_{norm} based on channel information h in line 6, and then the values λ_{avg} are computed to match as closely as possible with most inputs x from lines 7 to 15.

Accordingly, because the adversarial attack goes through channel h, the i^{th} element of the perturbation N be computed as $N_i = \frac{N_i^{adv}}{h_i}$ such that its dimension has the same dimension as N_i^{adv} after going through the channel model, for $i = 1, ..., p$, p is the dimension of the complex-valued (in-phase) inputs. Also, to meet the transmit power constraint P_{max} at the adversary, a scaling factor δ is added such that $N^{div} = -\delta \times N$, where $\delta = \frac{\sqrt{P}}{\|N\|_2}$. Thus, the perturbations (N) received in Equation (6) will be $N = N^{div} = -\delta \times N^{adv}$, where the minus indicates that the perturbation N is intended to decrease the recognizer's confidence in the true signals and increase its confidence in the target signal.

Algorithm 3: Semantic-oriented gradient descent attack in semantic communications

1 **Input:** Speech data set X, number of iteration α = 1000, perturbation budget ϵ = 0.001, step size λ =0.1.
Result: Adversarial perturbation N_{pgd}.
2 Initialize:$N_{pgd} = 0$
3 **for** *each iteration in range(α)* **do**
4 $\quad$ Calculate gradients in range of ϵ
5 $\quad grad = random.uniform(-\epsilon, \epsilon, size = data)$
6 $\quad$ Computed the perturbation as a $sign(.)$ of gradients.
7 $\quad \delta = sign(grad)$
8 $\quad$ Add perturbation into data
9 $\quad N_{pgd} = X + \epsilon\delta$
10 $\quad$ Clip N_{pgd} to ensure within valid range
11 $\quad N_{pgd} = fine - tuning(N_{pgd}, X - \epsilon, X + \epsilon)$
12 **end**
13 **Return** N_{pgd}

C. Semantic-oriented gradient descent attack

The projected gradient descent (PGD) attack method [46] is an iterative method that employs random initialization to generate adversarial examples. This method is more effective than the FGSM approach in increasing the impact of underfitting adversarial examples. This is achieved through the use of a linear approximation of the decision boundary around the data point. However, this process can lose some important features that limit the generated perturbations. The procedure for PGD is outlined in Algorithm 3. In this algorithm, the computation gradient and perturbation δ of each frame occurs in lines from 4 to 7. The noise addition and perturbation limitation are computed in lines 8–11, with a fine-tuning process to generate robust perturbations. This iterative process enhances attack efficiency, as detailed in Section VII. The tailored PGD in this work targets semantic-based speech applications.

VI. Semantic-non-targeted adversarial attacks and calibrations

The main goal of this section is to generate perturbation ϕ_k to misguide the model and mispredict any voice command, instead of specific commands as in the targeted attacks. Note that, to launch the targeted attack, three conditions must be met: the attacker knows (1) the exact input structure, (2) channel information between the transmitter and the receiver, and (3) the speech recognition model at the receiver. Those assumptions are not always practical for wireless channels. This section extends the attack capability by developing a novel technique that can mispredict any voice commands.

A. Universal adversarial perturbation attack with input-agnostic data in semantic communications

Adversarial perturbations are generated based on the input, with each input x requiring a corresponding perturbation to fool the model. This approach assumes knowledge of the model's input structure, which is often impractical,

such as knowing all supported commands. Input-agnostic methods address this by creating robust universal adversarial perturbations (UAPs) that fool the model regardless of input structure. In machine learning and computer vision, this technique is called a "universal adversarial perturbation" (UAP) [47]. It involves creating an adversarial sample in each iteration, making it computationally expensive. To address this, this study proposes a method to generate UAPs with lower complexity and improved fooling ability on speech datasets compared to the original UAPs. The fundamental rationale underlying the algorithm is as follows: Assuming that the subset $\{x_1, \ldots, x_k\} \in R$ is an arbitrary set of input, and their corresponding noises are subset $\{n_{x_1}, \ldots, n_{x_k}\}$, where $n_{x_i} = \bigtriangledown_{x_i} L(\alpha, x_i, y) / \| \bigtriangledown_{x_i} L(\alpha, x_i, y)\|_2$. The result of n_{x_i} is a noise matrix series of input subsets, where the first element in this matrix will have maximum variance. Instead of selecting a random element, this work chooses the largest in the perturbation set to maximize the model's fooling ability. The computation process is detailed in Algorithm 4, where it computes the first element v_1 which corresponds to the dominant right singular vector Ve_1 based on the first basis vector e_1 in line 4. This vector captures the most significant perturbation direction in the data. Then scale v_1 to the maximum allowed norm, P_{max}, to ensure the perturbation remains within the desired magnitude.

Algorithm 4: Universal adversarial perturbation attack with input-agnostic data

1 **Input:** A random set of input data $\{x_1, \ldots, x_k\}$, the model M, maximum allowed perturbation norm P_{max}
Result: UAP noise N_{uap}.
2 Initialize:$N_{uap} = 0$
3 Evaluate $X^{K \times p} = \{n_{x_1}, \ldots, n_{x_k}\}^T$
4 Compute the first element of $X = U \sum V^T$ and $v_1 = Ve_1$.
5 $N_{uap} = P_{max} v_1$
6 **Return** N_{uap}

B. Naive attack with lack of knowledge on channel information and the victim's model

Unlike the previous attack type, the objective of this method is to generate perturbations that do not have any channel information, e.g., the attacker has no knowledge about the victim's model [17]. They launch random attacks, resulting in random negative impacts on the semantic quality at the receiver. The attack process is outlined in Algorithm 5, which includes several phases. First, based on the power value P_{max}, the attacker splits it into the number of iterations in A and computes the loss function's gradient used to distort the transmitted signal on the channel. Second, the attacker computes the gradient again from the transmitter and adds perturbation. The attacker then adds another perturbation λ with $\frac{P_{\text{max}}}{A}$ into the signal (as shown in line 7 in Algorithm 5). This step increases the values in the loss function, causing important signals to be removed in the proposed system, thus increasing the signal distortion ability. The attacker then employs an iterative approach, repeating the method A times and this value is updated in each iteration.

Algorithm 5: Naive attack with lack of knowledge on channel information and the victim's model

1 **Input:** Input X, number of interaction $A = 1000$, power constraint P_{max}.
Result: Adversarial perturbation N_{naive}.
2 Initialize: Sum of gradient $\Delta = 0$, $X \rightarrow x$
3 **for** *i in range(A)* **do**
4 Calculate gradients in the range of A
5 $\lambda = \frac{\bigtriangledown_x L(\alpha, x, z)}{\| \bigtriangledown_x L(\alpha, x, z)\|_2}$
6 Compute other noise into input data
7 $x = x + \sqrt{\frac{P_{max}}{A}} h^{ch} \lambda$
8 Computed Δ
9 $\Delta = \Delta + \sqrt{\frac{P_{max}}{A}} \lambda$
10 **end**
11 $N_{naive} = \sqrt{P_{max}} \frac{\beta}{\|\beta\|_2}$
12 **Return** N_{naive}

C. DeepFool attack against multiple classifiers

The DeepFool attack method is a non-targeted attack that can be extended and implemented on general nonlinear classifiers and multi-class classifiers [48]. Its objective is to generate as smallest noise as possible by considering the minimizing distance between the clean (original) speech and the upper limitation of adversarial perturbations. The DeepFool method's perturbation, which is believed to be from a different input, can easily lead to model misclassification. As a result, this attack achieved high performance. In this study, we enhance this technique to target multiple classifiers in semantic communications, e.g., classifying spoken words into different commands such as "play," "pause," "stop," or identifying different spoken commands in multilingual speech data. The equation to generate a DeepFool adversarial example is computed by

$$\begin{aligned} N_{deepf} &= \underset{p_i}{\text{argmin}} \|p_i\|_2 \\ \text{subject to} \quad & f(x_i) + (\nabla f(x_i))^T p_i = 0, \end{aligned} \tag{18}$$

This algorithm stops when $sign(f(x_i)) \neq sign(f(x_0))$, where p_i is the i^{th} perturbation, the classifier f is the linear function, x_i is the i^{th} sample, x_0 is the clean sample.

VII. Evaluation results and attack performance

This section details the evaluation results of the attacks' performance on the semantic quality that is presented in Sections V and VI. The system transmits the neural network's output, which we consider as the semantic features. The quality of these semantic features is evaluated through PESQ's and SDR's scores as well as the feature's loss between before and after the attack. Therefore, we train 1000 speech, including 800 files to train and 200 files to test from the Edinburgh DataShare dataset [2], where set $M = 16.384$, sample frame

[2]https://datashare.ed.ac.uk/handle/10283/3061

TABLE IV
TOTAL PESQ'S SCORE UNDER MULTIPLE CHANNEL ATTACKS

Fading	Original score	Targeted attacks				Non-targeted attacks			
		FGSM	Channel inversion	PGD	Corresponding reduced score	UAP	DeepFool	Naive	Corresponding Reduced score
		(1)	(2)	(3)	(1)-(2)-(3)	(4)	(5)	(6)	(4)-(5)-(6)
AWGN	91.16	90.30	80.53	90.66	0.85 - <u>10.62</u> - 0.49	82.76	90.94	88.95	<u>8.39</u> - 0.21 - 2.20
Rayleigh	96.41	93.91	82.18	94.09	2.49 - <u>14.22</u> - 2.31	88.96	94.07	93.96	<u>7.45</u> - 2.33 - 2.45
Rician	104.28	96.04	82.78	95.77	8.24 - <u>21.50</u> - 8.51	96.31	100.74	96.53	<u>7.97</u> - 3.54 - 7.75

$f = 128$, and $l = 12$ is the length of each frame. All attacks are tested with a perturbation budget of $\epsilon = 0.001$, batch size = 32, learning rate 0.0001, and $\gamma = 1.0$ will target better [20] compared to other cases. The change in semantic quality corresponds to the change in PESQ and SDR scores under various attacks with $SNR = 8dB$. We only test the proposed system at $SNR = 8dB$ because it represents the most suitable value for the telephone semantic communication system [3]. The comparison clarifies which attack type causes the most signal distortion in a semantic communication system for speech data. Targeted (TA) and non-targeted (NTA) attacks are evaluated under identical settings.

A. Fundamental performance metrics for measuring the effectiveness of the proposed attacks

In our model, the system aims to evaluate semantic quality after channels are attacked by three metrics. First, this study uses the PESQ score to measure the semantic quality, as proposed by the ITU-T [49]. The voice signal quality threshold is set from -0.5 to 4.5 to meet the requirements for evaluating performance. Second, SDR [50] score is calculated to evaluate the change of score between the original speech x and the attacked speech $\overline{x}$, which is one of the commonly used metrics for speech transmission and can be expressed by

$$SDR = 10 * log_{10}\left[\frac{\|x\|^2}{\|x - \overline{x}\|^2}\right] \tag{19}$$

Based on this metric, if the SDR's value is higher on each channel, the voice information quality is better, i.e., human voice can be recognized more easily. The SDR metric, which stores the relative perturbation in relation to the noise, is used to assess the influence on semantic quality when comparing the changes in SDR scores on the channel. Additionally, the semantic quality of speech signals is intuitively manifested by a listening experience without latency and background noise, i.e., face-to-face conversation. Third, we also introduce a new metric, called data rate error (DRE) to evaluate the quality and reliability of semantic communication system. It reflects how efficiently data is transmitted and highlights potential issues like data loss or corruption. In this study, its computation formula is given by

$$DRE = \frac{\overline{\psi} - \psi}{\overline{\psi}}, \tag{20}$$

where ψ represents the received data and $\overline{\psi}$ represents the total data sent to receiver. In this metric, if $DRE = 0$, all data sent was received correctly (no error). If $DRE = 1$, no data was received (complete loss). If $0 < DRE < 1$, it indicates the fraction of data that was lost or not correctly received during transmission.

B. Perceptual evaluation of speech quality performance

Before the attack, PESQ values increase linearly in the range from $SNR = 0dB$ to $SNR = 20dB$. This demonstrates that the semantic quality is not affected and is an ideal environment for any wireless semantic communication system. However, this is not realistic because there are many potential risks that attackers want to explore to capture the victim's important information. Therefore, attacks on multiple channel models are deployed as to address challenges in this study. We simulate the results of those attacks as shown in Fig. 4. For targeted attacks, the PESQ scores are significantly lower compared to non-targeted attacks.

The FGSM and PGD attack methods perform less effectively, even worse than non-targeted attack methods when working on Rayleigh and Rician channels. We have listed these changes in PESQ scores in Table IV, where the most influenced results are highlighted in blue. Its PESQ score exhibits minimal variation compared to the original baseline (no attack). For non-target attacks, the PESQ scores for the DeepFool and UAP attacks uniformly decrease within a fixed range across all three channels across all SNR values, as illustrated in Table IV. However, the semantic features still increase linearly along with the increasing SNR. This demonstrates that the PESQ values change significantly, but the semantic quality remains smooth on the channel, i.e., the receiver does not perceive that the signal has been compromised compared with the original signal (no attack). As shown in Table IV, the UAP no-target attack significantly degraded semantic quality, reflected in the reduced PESQ score, due to the UAP algorithm's robustness and multi-directional attack capability.

Furthermore, the UAP method is deployed based on the PCA optimization processing, which also enhances the robustness of this attack method. Therefore, this attack method performs better compared to other attacks in terms of semantic quality and channel effects, while the other attack methods perform poorly for the majority of the SNR values. The obtained results are due to the non-targeted hostile attack's ability to move in any direction. For targeted attacks, we know that there are only ten distinct orientations because there are eleven corresponding modulation types [19]. However, the

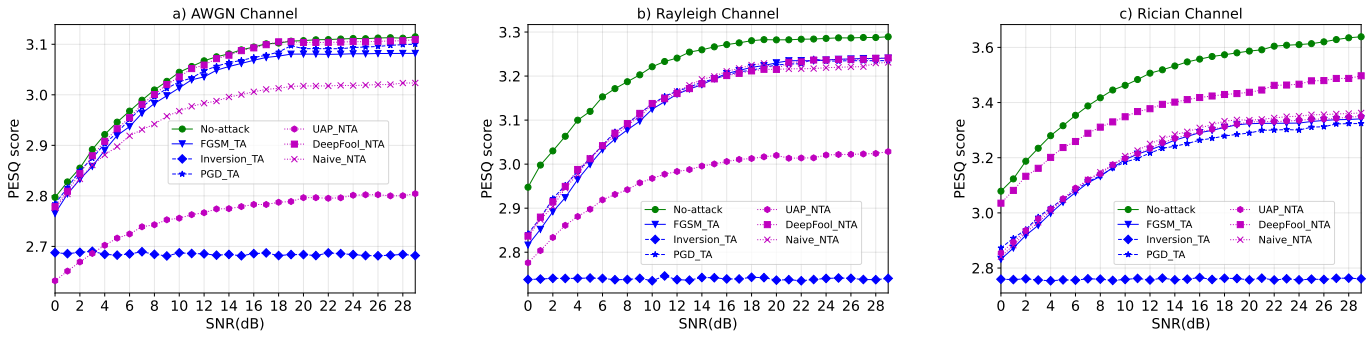

Fig. 4. PESQ score versus SNR on the proposed system under non-targeted attacks (pink lines) and targeted attacks (blue lines) for speech-based telephone communication. The average score ratio of channels AWGN, Rayleigh, and Rician is significantly reduced.

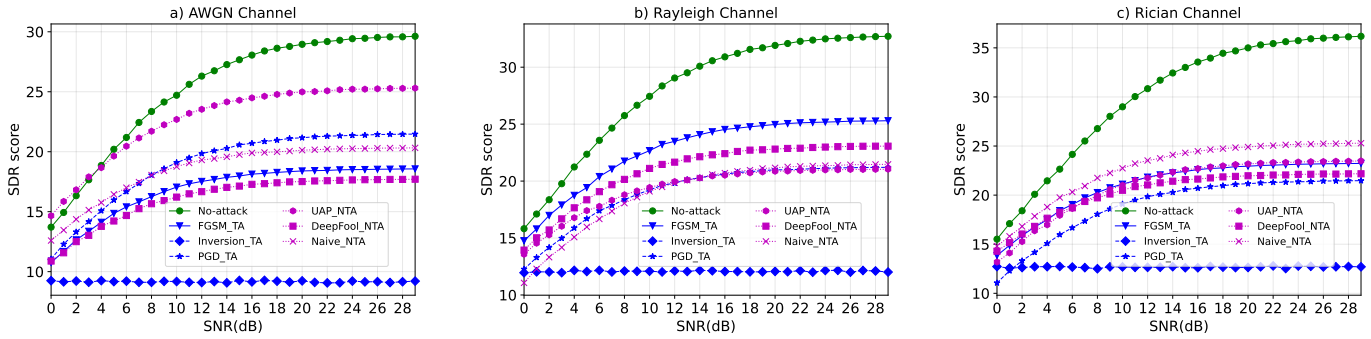

Fig. 5. The SDR score changes after attacking the proposed system under non-targeted attacks (pink lines) and targeted attacks (blue lines) versus SNR for speech-based telephone communications. The semantic quality in Rayleigh and Rician channels is most affected by targeted attacks, while non-targeted attacks primarily affect the AWGN and Rician channels.

TABLE V
TOTAL SDR SCORE UNDER MULTIPLE CHANNEL ATTACKS

Fading	Original score	Targeted attacks				Non-targeted attacks			
		FGSM (1)	Channel inversion (2)	PGD (3)	Corresponding reduced score (1)-(2)-(3)	UAP (4)	DeepFool (5)	Naive (6)	Corresponding Reduced score (4)-(5)-(6)
AWGN	758.69	681.38	274.92	569.12	77.31 - 483.77 - 189.57	504.93	483.77	556.96	253.76 - 274.92 201.73
Rayleigh	840.02	681.23	362.14	573.56	158.79 - 477.88 - 266.46	580.00	627.75	569.12	260.02 - 212.27 - 270.90
Rician	900.71	630.46	380.13	569.12	270.25 - 520.58 - 331.59	628.98	610.85	680.88	271.73 - 289.86 - 219.83

non-targeted attacks do not have such limitations. As a result, it is more likely that the non-targeted attack will choose a better way to force misrecognition, leading to changes in the series on the semantic quality in semantic communication systems. In addition, the computing complexity of non-targeted attacks is much lower than that of targeted attacks, which require iterations to achieve the desired accuracy. Finally, we observe that when the SNR level increases for the attacks, the semantic quality of the Rician channel decreases the most, followed by the Rayleigh and AWGN channels. Additionally, both TA and NTA attacks will only affect some specific areas from $SNR = 12$ to $SNR = 19$ on each channel, as shown in Fig. 4. The semantic quality in these areas will change the most. Based on the findings, we discovered that targeted assaults impair speech semantic quality for wireless systems in complex communication scenarios more than non-targeted attack techniques, particularly in the low SNR regime.

C. Signal-to-distortion ratio performance results

The SDR score is used to assess the semantic quality on each channel: a higher SDR score indicates that the semantic quality is better, i.e., the receiver can clearly receive the semantics in the signal. Fig. 5 shows the impacts of TAs and NTAs on the change in SDR scores under the AWGN, Rayleigh, and Rician channels, with the SNR score adjusted from $0dB$ to $19dB$. The figure shows that the targeted attacks achieved higher performance and reduced the SDR score more than the non-targeted attacks under all tested channel environments, while its performance remains reliable when SNR is high. The SDR score reduces more with channels that have higher fading, as shown in Fig. 5 b) and c). Also, the targeted attacks based on the obtained information, e.g., channel information or model parameters, always have high performance. Therefore, the adversarial attack performance achieved such differences. This is explained by the targeted

hostile attack's ability to move only in the targeted-selection direction instead of the free direction in non-targeted attacks.

However, the computational complexity for targeted attacks is higher compared to non-targeted attacks. The main reason is the iterations required to reach the desired accuracy. In our experiments, their semantic quality changes under semantic targeted attack are presented in Table V. The SDR score of targeted attacks on the AWGN channel is reduced to 1.35 times compared to non-targeted attacks. This rate on Rayleigh and Rician channels is 2.68 and 3.02 times, respectively. As a result, the semantic quality on the Rayleigh and Rician channels is mostly reduced under targeted attacks, followed by the AWGN channel. These changes mainly involve their propagation environment (i.e., with LOS/NLOS paths) on each corresponding channel. This also significantly impacts the semantic quality of telephone SEMCOM systems.

To mitigate the impact of these attacks and maintain the PESQ's and SDR's score in SEMCOM system, it is crucial to create tailored adaptive security systems for each communication channel. This will significantly minimize the vulnerabilities of the system. For example, the human can speak near the smart received device to avoid the strong manipulation of the attackers by injecting malicious signals into one of the NLOS paths. However, this may create inconvenience for users using their automatic features in smart homes and smart healthcare. Alternatively, encryption and authentication mechanisms can secure channel access, ensuring voice data integrity, confidentiality, and high-quality transmission.

D. Data rate error evaluation

DRE is used to evaluate the efficiency of a semantic communication system by indicating the amount of data errors over total transmitter data from the sender to the receiver. An attack causing higher DRE means more efficiency than that with lower DRE. The evaluation results in Fig. 6 show that the error rate of transmitted data on channel modeling is changed significantly among targeted and non-targeted attacks under various channels, i.e., comparison among AWGN, Rayleigh, and Rician channels. For non-targeted attacks, UAP and DeepFool both result in higher DRE values compared to the non-targeted attack scenario, especially at lower SNR values (near $0dB$). This indicates that these attacks increase data rate errors, making communication less reliable. The naive attack also causes elevated DRE values, but its performance appears to be comparable to or slightly less effective than DeepFool and UAP in disrupting the signal at lower SNR. In short, as SNR increases beyond $8dB$, the DRE values for all attacks (UAP, DeepFool, and Naive) tend to stabilize, indicating that at higher SNR, the effect of these attacks is less pronounced. For targeted attacks, all three attacks show significant increases in DRE from $0dB$ to $4dB$, with FGSM and inversion attack methods being more impactful (higher DRE values) than PGD. The DRE values for all attacks converge quickly as SNR rises, reaching similar levels to the no-attack case by around $6dB$. This also indicates that the synchronization of the noise signs with input speech only causes the signal obfuscation at SNR lower, i.e., SNR in the range $0dB$ to $8dB$. This could be a sign that your data has been compromised. For SNR values above $6dB$, the DRE values for all attack scenarios stabilize close to the No-attack DRE level, around 0.3×10^{-4}. This suggests that at higher SNRs, the attacks have minimal additional impact on DRE, and the system becomes more resilient to adversarial perturbations. The receiver makes it very difficult to detect the difference between input speech and noise added.

The channel inversion attack performs the best across all channels. As the results indicated in Figs. 4, 5, and 6, the attack's PESQ, SDR scores, and DRE are the lowest compared to other attack methods. This is because the attackers captured the distribution of the channel h_i between the sender and the receiver and attacked by applying the noise set $N_i = \frac{N_i^{adv}}{h_i}$ to channel h of the SEMCOM system. Channel information is fully utilized to generate robust perturbations N^{adv}. We found that non-targeted attacks ignoring channel effects perform poorly. This is because the wireless channel alters the phase and magnitude of perturbations at the receiver, with variations across channels affecting attackers' ability to capture information. This leads to signal degradation, differing in AWGN, Rayleigh, and Rician channels. In addition, the targeted channel inversion attack outperforms the non-targeted naive attack, which is limited by its lack of channel information. This indicates that the received power of perturbations significantly influences the performance of the classifier at the receiver during these attacks. The degradation rate of semantic quality depends on the channel information captured by the attacker. Thus, channel information is vital for data transmission, and its capture by an attacker poses security risks to the victim.

E. Loss function results

To demonstrate the efficiency of our method in maximizing perturbations on the loss function, we simulate the impact of noise on loss function under attacks, as shown in Fig. 7. This change is reflected in the relationship between the MSE loss values and the number of epochs. We can observe that the MSE loss changes significantly in both types of attacks. After approximately 400 epochs, the variation of these signals decreases with an $SNR = 8dB$. In any communication system, as this loss value approaches 0 after iterating through epochs, the accuracy of the system increases. In Fig. 7, The initial value of the system's loss function for the original signals (no attack) is closest to zero. After the attack, the loss values of the proposed system increasingly far from both the value of 0 and the initial system's loss values on all channels. This difference is illustrated by the red arrow in Fig. 7.

The loss value increases if the number of epochs increases. We observed that these loss values are increasing on each channel, with the Rician channel experiencing the greatest increase (i.e., its distance is largest from the value of 0). For semantic quality, it is evident that the Rician channel is more impacted compared to another channel with the same attack type, leading to a substantial degradation in semantic quality at the receiver. This also makes it difficult to find convergence between adversarial signals and original signals. This is a significant limitation of adversarial examples because if this convergence is achieved, it indicates that the attacker

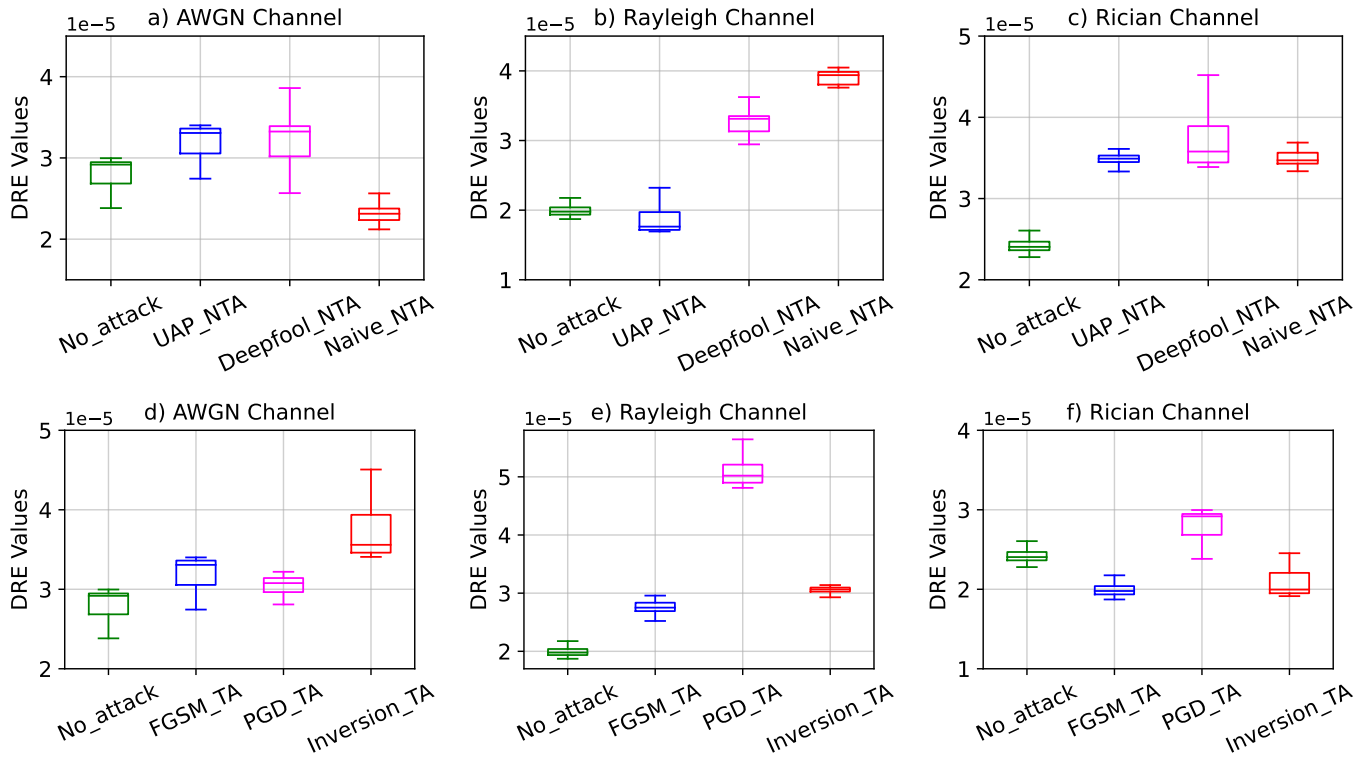

Fig. 6. The results of the data rate error values for the proposed system under targeted (bottom) and non-targeted (top) attacks. The targeted attacks cause the most signal degradation compared with original signals on AWGN and Rician channels. The data loss rate in target attacks calculated in bytes is up to 5% compared with non-target attacks. The reliability of the data under Rayleigh and Rician channels in targeted attacks is significantly lower compared to non-targeted attacks because their sample median is the closest to the minimum values.

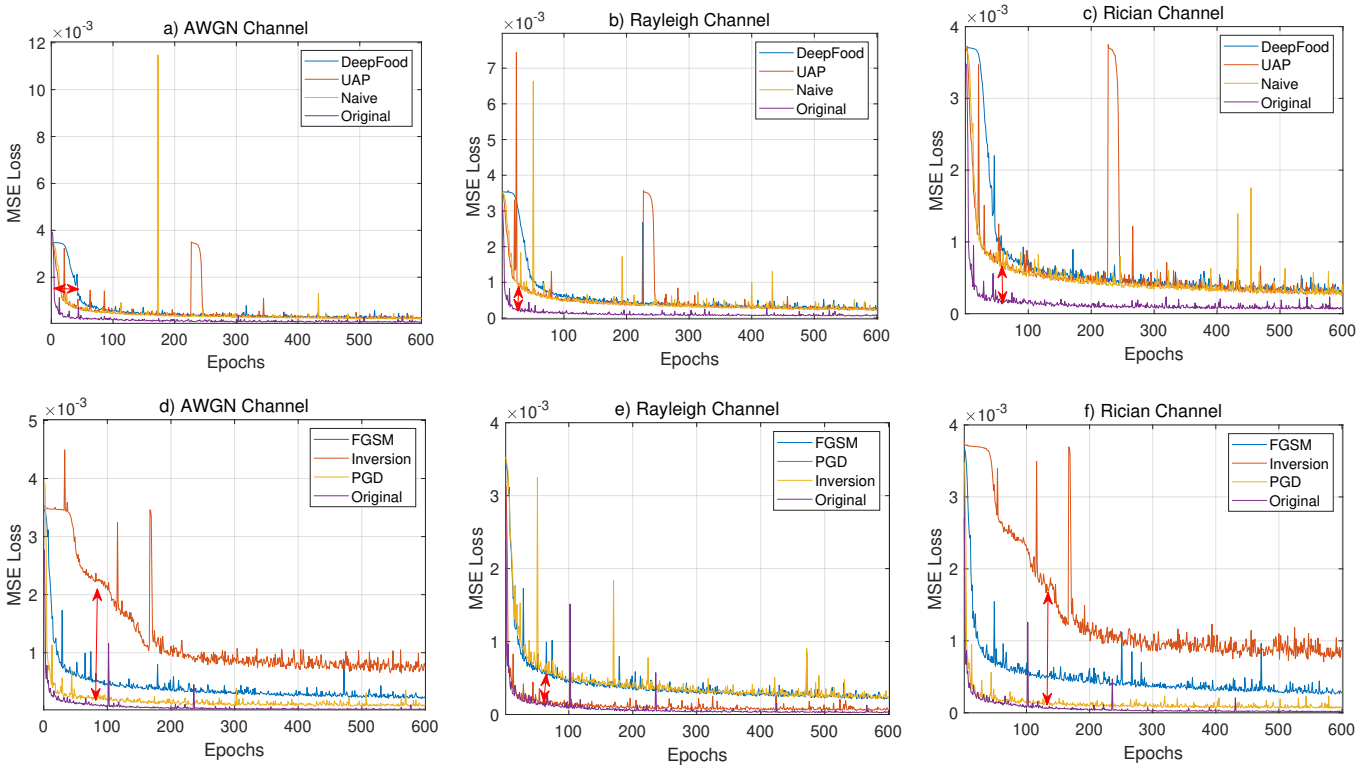

Fig. 7. The loss function attack results for the proposed system under targeted (bottom) and non-targeted (top) attacks. The left-side column is the AWGN channel, the middle column is the Rayleigh channel, and the right-side column is the Rician channel. Non-targeted attacks cause the most signal loss (i.e., signals are easily changed compared to targeted attacks) compared with original signals on AWGN and Rician channels. Double arrows show the loss evaluation under attacks compared with the original loss values over epochs.

has successfully crafted adversarial examples that can bypass the model's defenses and potentially cause extremely harmful outcomes, such as misclassification or incorrect decision-making. This also is an advantage of the non-targeted attacks in selecting the most efficient attack direction. Additionally, we found that the UAP attack increasingly impacts the system more significantly as the number of epochs increases, ranging from 200 to 300 epochs. This is due to the fact that the UAP attack algorithm is supported by the PCA algorithm during perturbation optimal processing. This impact will be reduced with other attacks. In the targeted attacks visualized at the bottom of Fig. 7, we observe that the loss mainly focuses on AWGN and Rician channels. The channel inversion attacks cause the most significant loss, followed by FGSM attacks, compared with non-targeted attacks, as indicated by the red double arrow. Therefore, the semantic features are significantly reduced under targeted attacks on both channels. For SHAP

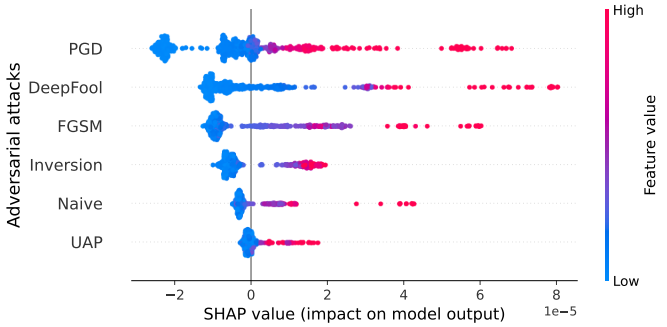

Fig. 8. The changes of the SHAP value occur during the entire speech training of the system under all attacks. Each dot in the plots represents a semantic feature. The colors used indicate the magnitude of the semantic feature value (red representing important values and blue representing less important values). Finally, the position on the horizontal axis represents the impact of the semantic feature value on the prediction of the target.

values, Fig. 8 presents a summary of Shapley plots for all adversarial attacks on the system. This summary plot illustrates the positive and negative relationships with the quality of semantics, which are impacted after attacks. We notice that the most crucial semantic features exhibit higher SHAP values. This suggests that these features are more attack-ability and prone to easy changes after each training epoch. Overall, the markers red dot has the greatest influence on the decision-making of our model in recognition semantics in the telephone semantic communication systems. If these values are altered through an attack, the semantic quality will subsequently change. In addition to training the model with a larger set of adversarial examples to mitigate attack performance, we can also combine the properties of MSE and MAE (Mean Absolute Error) in the speech signal reconstruction process to make the system less sensitive to residual.

F. Fine-tuning the sensitivity to residuals for adversarial defense in wireless SEMCOM

The defense mechanism can be achieved partially with the desired confidence by using selected smoothing in the sensitive SNR ranges, i.e., ranges $2dB - 6dB$ and $12dB - 16dB$ in

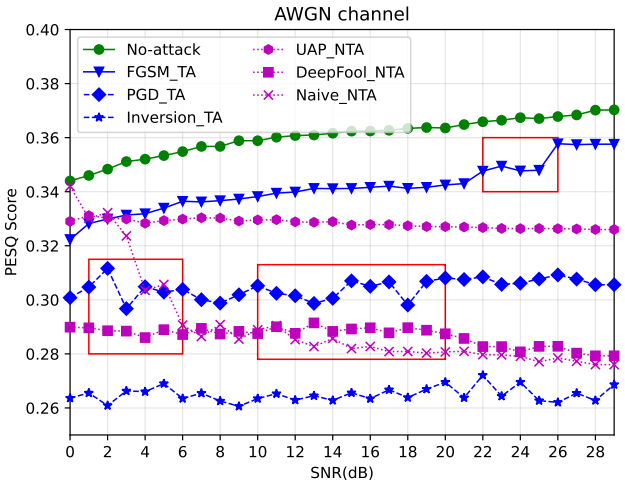

Fig. 9. Fine-tuning by removing the sensitive noise ranges (range marked by red) of weights on AWGN transmission channel on SNR's value. The final target is to make smooth for weights on the transmission channel.

Fig. 9. Based on this range, we can apply the fine-tuning method to achieve a smooth range. For example, we can remove the residual value $L_\delta(a)$ according to Huber loss as in Equation (21) below, where $a = x - \overline{y}$ (difference between the x actual and $\overline{y}$ predicted values) based on a threshold δ which is identified by the user's system.

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta \\ \delta(|a| - \frac{1}{2}\delta) & \text{for } |a| > \delta \end{cases} \tag{21}$$

However, this method is not the perfect approach either because selecting the threshold δ to remove the residual values is also challenging. If we choose a value that is too large, important features will be eliminated. Conversely, if we choose a value that is too small, the redundant values will not be effectively removed as desired. In our study, selecting and identifying the $a = \phi_k$ to apply on Equation (21) is challenging. Because the perturbation signals are synchronized and trained with the input signals. The potential approach is to fix the weight range for sensitive noise signals based on the SNR's metrics as in Fig. 9. Then we can remove the mutant values in these ranges to give the smooth values for the channel. We can consider that the removed values are noise signals from attackers. The values without being removed still contain the perturbation, however, this noise proportion will not be significant because the clear input samples will occupy a larger proportion. As a result, the misclassification/misrecognition rate of the AI-enabled models will be almost zero.

VIII. Conclusion

As the implementation of deep learning models in semantic communication systems becomes more prevalent, it is vital to consider these applications from a security and robustness perspective. This is the first work to highlight the potential risks of wireless-based semantic communication systems. We simulate two types of physical-layer attacks against the semantic communication system that cause semantic quality changes at the receiver on AWGN, Rayleigh, and Rician channel

models. Additionally, we evaluate adversarial attacks to assess the semantic quality impact after attacking. Experiment results show that both targeted and non-targeted attacks significantly affect semantics at the receiver. Overall, the Rayleigh is fading, and Rician channels are more impacted (i.e., their score is reduced) than AWGN when there is prior information about the system. Furthermore, the reliability of the data under Rayleigh and Rician channels in targeted attacks is also significantly lower, by about 2.9 times, compared to the original input data (i.e., their sample median is the closest to the minimum values). This demonstrates that CNNs used for modulation classification are vulnerable to adversarial attacks in the low SNR regime. In addition, we discuss some of the specific defense methods to mitigate risks from adversarial attacks. Future research needs to focus on hardening SEMCOM systems using psychoacoustic models or previously authenticated voices to prevent attacks. The attacks should be also tested on commercial SEMCOM systems, e.g., machine-to-machine communications in smart factories, to enhance semantic quality on wireless channels. Given that these communications are likely to become key technologies in 6G intelligent networks, developing novel and efficient defense techniques represents a highly promising research area.

REFERENCES

[1] H. Xie and Z. Qin, "A lite distributed semantic communication system for internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 142–153, 2020.

[2] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge learning for b5g networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE journal of selected topics in signal processing*, vol. 17, no. 1, pp. 9–39, 2023.

[3] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2434–2444, 2021.

[4] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.

[5] R. Sun, Y. Zhang, T. Shah, J. Sun, S. Zhang, W. Li, H. Duan, B. Wei, and R. Ranjan, "From sora what we can see: A survey of text-to-video generation," 2024.

[6] S. Zhao, Y. Zhang, X. Cun, S. Yang, M. Niu, X. Li, W. Hu, and Y. Shan, "Cv-vae: A compatible video vae for latent generative video models," 2024.

[7] D. Huang, X. Tao, F. Gao, and J. Lu, "Deep learning-based image semantic coding for semantic communications," in *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, IEEE, 2021.

[8] C. Dong, H. Liang, X. Xu, S. Han, B. Wang, and P. Zhang, "Innovative semantic communication system," *arXiv preprint arXiv:2202.09595*, 2022.

[9] D. Adesina, C.-C. Hsieh, Y. E. Sagduyu, and L. Qian, "Adversarial machine learning in wireless communications using rf data: A review," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 77–100, 2022.

[10] V.-T. Hoang, Y. A. Ergu, V.-L. Nguyen, and R.-G. Chang, "Security risks and countermeasures of adversarial attacks on ai-driven applications in 6g networks: A survey," *Journal of Network and Computer Applications*, p. 104031, 2024.

[11] Y. E. Sagduyu, T. Erpek, S. Ulukus, and A. Yener, "Is semantic communication secure? a tale of multi-domain adversarial attacks," *IEEE Communications Magazine*, vol. 61, no. 11, pp. 50–55, 2023.

[12] F. Jiang, L. Dong, Y. Peng, K. Wang, K. Yang, C. Pan, and X. You, "Large ai model empowered multimodal semantic communications," *IEEE Communications Magazine*, pp. 1–7, 2024.

[13] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro, "Intriguing properties of adversarial ml attacks in the problem space," in *2020 IEEE symposium on security and privacy (SP)*, pp. 1332–1349, IEEE, 2020.

[14] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Transactions on Signal Processing*, vol. 69, pp. 2663–2675, 2021.

[15] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Wireless image retrieval at the edge," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 1, pp. 89–100, 2020.

[16] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Joint device-edge inference over wireless links with pruning," in *2020 IEEE 21st international workshop on signal processing advances in wireless communications (SPAWC)*, pp. 1–5, IEEE, 2020.

[17] G. Nan, Z. Li, J. Zhai, Q. Cui, G. Chen, X. Du, X. Zhang, X. Tao, Z. Han, and T. Q. Quek, "Physical-layer adversarial robustness for deep learning-based semantic communications," *IEEE Journal on Selected Areas in Communications*, 2024.

[18] X. Peng, Z. Qin, D. Huang, X. Tao, J. Lu, G. Liu, and C. Pan, "A robust deep learning enabled semantic communication system for text," in *GLOBECOM 2022-2022 IEEE Global Communications Conference*, pp. 2704–2709, IEEE, 2022.

[19] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Channel-aware adversarial attacks against deep learning-based wireless signal classifiers," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 3868–3880, 2021.

[20] B. Kim, Y. E. Sagduyu, K. Davaslioglu, T. Erpek, and S. Ulukus, "Over-the-air adversarial attacks on deep learning based modulation classifier over wireless channels," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, IEEE, 2020.

[21] M. Sadeghi and E. G. Larsson, "Adversarial attacks on deep-learning based radio signal classification," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 213–216, 2018.

[22] Y. Huang, X. Li, W. Wang, T. Jiang, and Q. Zhang, "Forgery attack detection in surveillance video streams using wi-fi channel state information," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4340–4349, 2021.

[23] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *ICC 2021-IEEE International Conference on Communications*, pp. 1–6, IEEE, 2021.

[24] R. Schulz, O. Günlü, R. Elschner, R. F. Schaefer, C. Schmidt-Langhorst, C. Schubert, and R. F. Fischer, "Semantic security for indoor thz-wireless communication," in *2021 17th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 1–6, IEEE, 2021.

[25] Z. Ahmad, S. J. Hashim, G. Ferre, F. Z. Rokhani, S. Al-Haddad, and A. Sali, "Lora and rotating polarization wave: Physical layer principles and performance evaluation," *IEEE Access*, vol. 11, pp. 14892–14905, 2023.

[26] Y. E. Sagduyu, Y. Shi, T. Erpek, W. Headley, B. Flowers, G. Stantchev, and Z. Lu, "When wireless security meets machine learning: Motivation, challenges, and research directions," *arXiv preprint arXiv:2001.08883*, 2020.

[27] F. Wang, C. Zhong, M. C. Gursoy, and S. Velipasalar, "Adversarial jamming attacks and defense strategies via adaptive deep reinforcement learning," *arXiv preprint arXiv:2007.06055*, 2020.

[28] D. Ma, Y. Wang, and S. Wu, "Against jamming attack in wireless communication networks: A reinforcement learning approach," *Electronics*, vol. 13, no. 7, p. 1209, 2024.

[29] J. Lin, *Adversarial and data poisoning attacks against deep learning*. PhD thesis, University of South Florida, 2022.

[30] Z. Luo, S. Zhao, Z. Lu, J. Xu, and Y. E. Sagduyu, "When attackers meet ai: Learning-empowered attacks in cooperative spectrum sensing," *IEEE Transactions on Mobile Computing*, vol. 21, no. 5, pp. 1892–1908, 2020.

[31] Z. Luo, S. Zhao, R. Duan, Z. Lu, Y. E. Sagduyu, and J. Xu, "Low-cost influence-limiting defense against adversarial machine learning attacks in cooperative spectrum sensing," in *Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*, pp. 55–60, 2021.

[32] Z. Luo, S. Zhao, Z. Lu, Y. E. Sagduyu, and J. Xu, "Adversarial machine learning based partial-model attack in iot," in *Proceedings of the 2nd ACM workshop on wireless security and machine learning*, pp. 13–18, 2020.

[33] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Over-the-air membership inference attacks as privacy threats for deep learning-based wireless signal classifiers," in *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, pp. 61–66, 2020.

[34] Y. Shi and Y. E. Sagduyu, "Membership inference attack and defense for wireless signal classifiers with deep learning," *IEEE Transactions on Mobile Computing*, vol. 22, no. 7, pp. 4032–4043, 2022.

[35] Y. Shi, K. Davaslioglu, and Y. E. Sagduyu, "Generative adversarial network in the air: Deep adversarial learning for wireless signal spoofing," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 1, pp. 294–303, 2020.
[36] Y. E. Sagduyu, T. Erpek, and Y. Shi, "Adversarial machine learning for 5g communications security," *Game Theory and Machine Learning for Cyber Security*, pp. 270–288, 2021.
[37] Y. Shi and Y. E. Sagduyu, "Adversarial machine learning for flooding attacks on 5g radio access network slicing," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, IEEE, 2021.
[38] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6g," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 140–145, 2022.
[39] Z. Li, J. Zhou, G. Nan, Z. Li, Q. Cui, and X. Tao, "Sembat: Physical layer black-box adversarial attacks for deep learning-based semantic communication systems," in *2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall)*, pp. 1–5, IEEE, 2022.
[40] J. Dai, P. Zhang, K. Niu, S. Wang, Z. Si, and X. Qin, "Communication beyond transmitting bits: Semantics-guided source and channel coding," *IEEE Wireless Communications*, vol. 30, no. 4, pp. 170–177, 2023.
[41] M. Sadeghi and E. G. Larsson, "Physical adversarial attacks against end-to-end autoencoder communication systems," *IEEE Communications Letters*, vol. 23, no. 5, pp. 847–850, 2019.
[42] Z. Yu, Y. Xiong, K. He, S. Huang, Y. Zhao, and J. Gu, "Position-invariant adversarial attacks on neural modulation recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3483–3487, IEEE, 2022.
[43] T. Zhang, "Faster augmented lagrangian method for solving convex optimization problems with linear constraints," 2024.
[44] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, 2016.
[45] J. S. Goodfellow, I.J.and Shlens, "Explaining and harnessing adversarial examples," *Machine Learning*, 2014.
[46] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.
[47] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.
[48] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," 2016.
[49] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.
[50] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

Van-Tam Hoang is currently pursuing a Ph.D. degree at the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. He is working with the Cyber Information Security Lab. His research interests include cybersecurity, deep learning, and semantic communications.

Van-Linh Nguyen (Senior Member, IEEE) is an Assistant Professor at the Department of Computer Science and Information Engineering, National Chung Cheng University (CCU), Taiwan, and the lead of the Cyber Information Security Laboratory (CIS Lab). He was also a postdoctoral fellow with CCU from 2020 to 2022. He received his Ph.D. in computer science and information engineering from CCU, in 2019. He has actively served as a reviewer for flagship TVT, COMMAG, COMST, COMML, and participated as a Technical Program Committee Member for a variety of international conferences, such as CISC 2023, ICTA 2023, and CITA 2024. He is a guest editor of The Internet of Military Defense Things special issue in IEEE Internet Of Things Magazine. His research interests include physical layer security, vehicular security, quantum machine learning, and wireless communications.

Rong-Guey Chang received his B.S. and M.S. degrees from National Chung Hsing University, Taiwan, in 1991 and 1993, respectively, and his Ph.D. degree from National Tsing Hua University, Taiwan, in 2000. He was an assistant professor from 2002 to 2009 and an associate professor from 2009 to 2014 at National Chung Cheng University, Taiwan. Since 2014, he has been a professor at National Chung Cheng University, Taiwan. His research interests are in robot, compiler, and embedded systems.

Po-Ching Lin (Member, IEEE) received his Ph.D. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 2008. He joined the faculty of the Department of Computer Science and Information Engineering, CCU, in August 2009. He is currently a professor. He serves as an associate editor of IEEE Communications Surveys & Tutorials. His research interests include network security, network traffic analysis, and performance evaluation of network systems.

Ren-Hung Hwang (Senior Member, IEEE) received his Ph.D. degree in computer science from the University of Massachusetts, Amherst, Massachusetts, USA, in 1993. He is the Dean of the College of Artificial Intelligence, National Yang Ming Chiao Tung University (NYCU), Taiwan. Before joining NYCU, he was with National Chung Cheng University, Taiwan, from 1993 to 2022. He is currently on the editorial boards of IEEE Communications Surveys & Tutorials, IEEE Transactions on Vehicular Technology, and IEICE Transactions on Communications. He received the Best Paper Award from The 6th International Conference on Internet of Vehicles 2019, IEEE UbiMedia 2018, IEEE SC2 2017, IEEE IUCC 2014, and the IEEE Outstanding Paper Award from IEEE IC/ATC/ICA3PP 2012. He served as the general chair of the International Computer Symposium (ICS), 2016, and International Symposium on Pervasive Systems, Algorithms, and Networks (ISPAN) 2018, International Symposium on Computer, Consumer and Control (IS3C) 2018, IEEE DataCom 2019 (The 5th IEEE International Conference on Big Data Intelligence and Computing). He received the Outstanding Technical Achievement Award from the IEEE Tainan Section in 2022, as well as the Outstanding Innovative Research Award and the Outstanding Industry-Academic Contribution Award from TACC, Taiwan, in 2023 and 2024, respectively. His research interests include deep learning, network security, wireless communications, Internet of Things, cloud and edge computing.

Trung Q. Duong (Fellow, IEEE) is a Canada Excellence Research Chair (CERC) and a Full Professor at Memorial University, Canada. He is also an adjunct professor at Queen's University Belfast, UK. He was a Distinguished Advisory Professor at Inje University, South Korea (2017-2019), an adjunct professor at Duy Tan University, Vietnam (2012-present), and a Visiting Professor (under Eminent Scholar program) at Kyung Hee University, South Korea (2023-2025). His current research interests include wireless communications, quantum machine learning, and quantum optimisation.

Dr. Duong has served as an Editor/Guest Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS LETTERS, IEEE WIRELESS COMMUNICATIONS, IEEE COMMUNICATIONS MAGAZINES, and IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He received the Best Paper Award at the IEEE VTC-Spring 2013, IEEE ICC 2014, IEEE GLOBECOM 2016, 2019, 2022, IEEE DSP 2017, IWCMC 2019, 2023, 2024 and IEEE CAMAD 2023, 2024. He is the Editor in Chief of IEEE Communications Surveys & Tutorials. He has received the two prestigious awards from the Royal Academy of Engineering (RAEng): RAEng Research Chair and the RAEng Research Fellow. He is the recipient of the prestigious Newton Prize 2017. He is the recipient of the prestigious Newton Prize 2017. He is a Fellow of the Engineering Institute of Canada (EIC) and Asia-Pacific Artificial Intelligence Association (AAIA).