Robust Semantic Communications for Speech Transmission

Zhenzi Weng* and Zhijin Qin†
* Department of Science and Engineering, Queen Mary University of London, London, UK
† Department of Electronic Engineering, Tsinghua University, Beijing, China
Email: zhenzi.weng@qmul.ac.uk, qinzhijin@tsinghua.edu.cn

***Abstract*—In this paper, we propose a robust semantic communication system for speech transmission, named Ross-S2T, by delivering the essential semantic information. Particularly, we consider the speech-to-text translation (S2TT) as the transmission goal. First, a deep semantic encoder is developed to directly convert speech in the source language to textual features associated with the target language, facilitating the end-to-end (E2E) semantic exchange to perform the S2TT task and reducing the transmission data without performance degradation. To mitigate semantic impairments inherent in the corrupted speech, a novel generative adversarial network (GAN)-enabled deep semantic compensator is established to estimate the lost semantic information within the speech and extract deep semantic features simultaneously, which enables robust semantic transmission for corrupted speech. Furthermore, a semantic probe-aided compensator is devised to enhance the semantic fidelity of recovered semantic features and improve the understandability of the target text. According to simulation results, the proposed Ross-S2T exhibits superior S2TT performance compared to conventional approaches and high robustness against semantic impairments.**

***Index Terms*—Deep learning, generative adversarial network, semantic communications, speech-to-text translation.**

I. Introduction

Semantic communications have been regarded as a promising solution to tackle the technical challenges in conventional communication systems and have attracted significant research attention in recent years [1]. The advancement of semantic communications derives from the ability to explore semantic information and achieve semantic exchange, which revolutionizes many aspects of wireless communications [2].

According to the three-level communication architecture proposed by Shannon and Weaver [3], semantic communications are the second level of communications that prioritize conveying the underlying meaning by representing the input message with minimal ambiguity through semantics, which overcomes the limitation of conventional communications to process data at the bit level and drives the evolution of intelligent communications. However, the investigation of semantic communication is in its infancy and there is no consensus on the definition of semantics, hindering the motivation to represent the semantic information with a rigorous formula. Inspired by the thriving of artificial intelligence (AI) in diverse areas, deep learning (DL)-enabled semantic communication paradigm breaks the bottleneck of a mathematical theory to quantify the semantics and has shown its great potential to learn semantic information by designing sophisticated neural networks (NNs). DL-enabled semantic communications have experienced unprecedented developments due to the ubiquity of intelligent mobile devices and the booming demand for semantic-driven data transmission.

According to the transmission goal, semantic communication systems can restore the source message and/or perform the AI tasks. The global semantic features are transmitted to achieve the data reconstruction. However, in the task-oriented semantic transmission mechanism, only the task-related semantic features are extracted, serving diverse user requests and enabling efficient transformation between multimodal data. Particularly, the pioneering work on DL-enabled semantic communications, named DeepSC [4], has been proposed to recover accurate text by leveraging the transformer module to learn textual semantic features. A variant of DeepSC, named R-DeepSC [5], has been devised to eliminate semantic noise and facilitate robust text transmission. In [6], Weng *et al.* introduced a semantic communication system for speech transmission, named DeepSC-S, which utilizes a joint semantic-channel coding scheme to extract and transmit global semantic features. Inspired by DeepSC-S, a deep speech semantic transmission scheme has been developed in [7] by adopting a flexible rate-distortion trade-off to achieve end-to-end (E2E) optimization. Huang *et al.* [8] considered a semantic communication system for image transmission and Jiang *et al.* [9] presented a video semantic transmission paradigm to alleviate semantic errors by incorporating a semantic detector.

Furthermore, Xie *et al.* [10] established a task-oriented semantic communications system for machine translation and visual question-answering tasks by fusing textual and visual semantic features. The speech recognition and speech synthesis tasks are performed in an efficient speech semantic transmission scheme by converting the speech input into the task-related semantic features [11]. In [12], Xu *et al.* proposed reinforcement learning-enabled semantic communications for scene classification in unmanned aerial systems. Zhang *et al.* investigated a semantic communication system for extended reality (XR) tasks by transmitting highly compressed semantic information to reduce network traffic. In [13], A unified multimodal multi-task semantic communication architecture,

The work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant62293484 and Grant 61925105; in part by the Beijing National Research Center for Information Science and Technology (BNRist), Beijing, China; and in part by the State Key Laboratory of Space Network and Communications, Beijing, China.

named U-DeepSC, has been developed by sharing trainable parameters amongst various tasks to reduce the redundancy of semantic features and accelerate the inference process.

In this paper, a robust semantic communication system for speech transmission, named Ross-S2T, is proposed. We argue that existing research works on task-oriented semantic communications for speech only extract textual semantic features constrained to the source language, i.e., shallow semantic features, encouraging the exploration of deep semantic features spanning various languages. Moreover, the intractable semantic impairments inherent in the corrupted speech are investigated. In this context, the speech-to-text translation (S2TT) task is considered in semantic transmission scenarios with corrupted speech input. The contributions of this paper are summarized as follows:

- A semantic communication system for S2TT in the context of clear speech, named DeepSC-S2T, is developed by utilizing a deep semantic encoder to extract textual semantic features related to the target language from speech in the source language.
- According to our comprehensive literature review, the speech semantic impairments in semantic communications are not investigated. We propose a generative adversarial network (GAN)-enabled deep semantic compensator to estimate the damaged semantic information inherent in the corrupted speech and generate deep semantic features simultaneously.
- To further reduce the semantic loss at the recovered features, a semantic impairment probe-aided compensator is established to perceive and calibrate the corrupted semantic features at the receiver, thereby improving the accuracy of the produced target text.
- Simulation results verify the superiority of the DeepSC-S2T to serve the S2TT task and the robustness of the Ross-S2T to contend with semantic impairments.

The rest of this paper is structured as follows. In Section II, the model of the robust semantic communications for S2TT is provided. Section III introduces the details of the proposed Ross-S2T. Section IV presents experimental results and Section V conclusions this paper.

II. System Model

In this section, we introduce robust semantic communication systems for speech transmission and consider S2TT as the transmission goal. The considered system aims to address two primary challenges. The first challenge is to deliver E2E semantic exchange and achieve efficient transmission from speech in the source language to text in the target language. The second one is to devise a semantic impairment suppression mechanism to contend with semantic impairments within the corrupted speech. To this end, the novel deep semantic codec mechanism is established to facilitate speech transmission for S2TT, and the deep semantic compensator is first developed to compensate for the complicated semantic impairments.

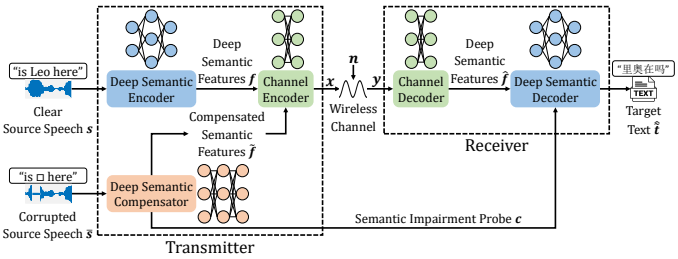


Fig. 1: The model structure of robust semantic communications for speech-to-text translation.

A. Clear Speech Input

The designed system is tailored for communication scenarios involving transmitter and receiver users from different linguistic backgrounds and facilitates the conversion of multimodal data from speech to text. The model structure of robust semantic communications for S2TT is shown in Fig. 1. From the figure, when the system input is clear speech, the deep semantic encoder compresses the speech sequence, $\boldsymbol{s}$, and extracts the deep semantic features, $\boldsymbol{f}$. Note that the deep semantic encoder directly learns the textual semantics in the target language. $\boldsymbol{f}$ is mapped to the symbols, $\boldsymbol{x}$, after passing through the NN-enabled channel encoder prior to transmission over the wireless channel. Then $\boldsymbol{x}$ can be expressed according to $\boldsymbol{s}$ as follows,

$$\boldsymbol{x} = \mathfrak{T}_{\mathrm{C}}(\mathfrak{T}_{\mathrm{S}}(\boldsymbol{s})) \quad \text{w.r.t.} \quad \boldsymbol{\alpha}, \tag{1}$$

where $\mathfrak{T}_{\mathrm{S}}(\cdot)$ and $\mathfrak{T}_{\mathrm{C}}(\cdot)$ are the deep semantic encoder and the channel encoder, respectively. $\boldsymbol{\alpha}$ is all trainable NN parameters of $\mathfrak{T}_{\mathrm{S}}(\cdot)$ and $\mathfrak{T}_{\mathrm{C}}(\cdot)$.

The encoded symbols, $\boldsymbol{x}$, are affected by the channel fading and channel noise after passing through the wireless channel layer, and the received symbols, $\boldsymbol{y}$, can be denoted as

$$\boldsymbol{y} = \boldsymbol{h} * \boldsymbol{x} + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{h}$ represents the fading channel and $\boldsymbol{n}$ is the additive white Gaussian noise (AWGN).

At the receiver, the symbols, $\boldsymbol{y}$, are recovered to the deep semantic features, $\widehat{\boldsymbol{f}}$, by the channel decoder. Moreover, the target text, $\widehat{\widehat{\boldsymbol{t}}}$, is obtained by feeding $\widehat{\boldsymbol{f}}$ into the deep semantic decoder, which can be modelled as

$$\widehat{\widehat{\boldsymbol{t}}} = \mathfrak{T}_{\mathrm{S}}^{-1}(\mathfrak{T}_{\mathrm{C}}^{-1}(\boldsymbol{y})) \quad \text{w.r.t.} \quad \boldsymbol{\beta}, \tag{3}$$

where $\mathfrak{T}_{\mathrm{C}}^{-1}(\cdot)$ and $\mathfrak{T}_{\mathrm{S}}^{-1}(\cdot)$ denotes the channel decoder and the deep semantic decoder, respectively. $\boldsymbol{\beta}$ represents all trainable NN parameters of $\mathfrak{T}_{\mathrm{C}}^{-1}(\cdot)$ and $\mathfrak{T}_{\mathrm{S}}^{-1}(\cdot)$.

B. Corrupted Speech Input

In practical communication systems, the integrity of input speech is highly susceptible to perturbations induced by the surrounding environments or the unstable network connection, thereby resulting in the potential degradation of original speech. In this work, our endeavours towards semantic communications for S2TT extend to the corrupted speech

input, wherein some semantic information within the speech is inaccessible due to the introduction of semantic impairments. As shown in Fig. 1, the corrupted speech, $\overline{\boldsymbol{s}}$, contains limited semantic information. Particularly, the deep semantic compensator tasks $\overline{\boldsymbol{s}}$ as input and generates the compensated deep semantic features, $\widetilde{\boldsymbol{f}}$, which predicts the lost information and extracts textual semantic features in the target language simultaneously, written as

$$\widetilde{\boldsymbol{f}} = \mathfrak{T}_{\mathrm{SC}}(\overline{\boldsymbol{s}}) \quad \text{w.r.t.} \quad \boldsymbol{\gamma}, \tag{4}$$

where $\mathfrak{T}_{\mathrm{SC}}(\cdot)$ indicates the deep semantic compensator and $\boldsymbol{\gamma}$ is the corresponding trainable NN parameters.

Furthermore, a semantic impairment probe, $\boldsymbol{c}$, containing an index vector with position information of corrupted semantic features, is attained and transmitted to the receiver over a reliable channel. The motivation behind the semantic impairment probe is to strengthen the semantic fidelity of the recovered deep semantic features by further reducing the semantic ambiguity caused by corrupted semantic features. It is noteworthy that the channel encoder, channel decoder, and deep semantic decoder trained in the context of clear speech input are shared in communication scenarios with corrupted speech input to generate the target text.

III. Robust Semantic Communications for Speech Transmission

To address the preceding challenges, we adopt a two-stage training scheme. Particularly, a semantic transmission paradigm for S2TT based on clear speech input, named DeepSC-S2T, is first proposed. Then, a dual-compensator mechanism is carried out to enhance robust semantic communications, named Ross-S2T, which utilizes a GAN-enabled deep semantic compensator and a semantic impairment probe-aided compensator to acquire as accurate deep semantic features as possible at the receiver.

A. DeepSC-S2T

The proposed Ross-S2T is shown in Fig 2. From the figure, at the first training stage, the convolutional neural network (CNN) module condenses the clear speech and the transformer module further extracts the features, $\boldsymbol{F}$, before passing through the dense layer-enabled channel encoder to attain symbols, $\boldsymbol{X}$. The dense layer constructs the channel decoder to process the receiver symbols, $\boldsymbol{Y}$, and the transformer-enabled deep semantic decoder is leveraged to produce multiple target text sequences, $\widehat{\widetilde{\boldsymbol{T}}}$. To boost the efficient semantic transmission for serving the S2TT task, the label-smoothing regularization-aided cross-entropy (LSR-CE) is adopted as the E2E loss function to train the DeepSC-S2T, which is expressed as

$$\mathcal{L}_{\mathrm{LSR-CE}}(\widetilde{\boldsymbol{T}},\ \widehat{\widetilde{\boldsymbol{T}}};\ \boldsymbol{\theta}) = \kappa \mathcal{L}_{\mathrm{CE}} + \sum_{\widetilde{l}=1}^{\widetilde{L}} f_w\left(w_e\right), \tag{5}$$

where $\kappa \in [0,1]$ is a hyperparameter that signifies the confidence level associated with the predicted tokens in $\widehat{\widetilde{\boldsymbol{T}}}$ matching the true tokens in the accurate text sequence in the target language, $\widetilde{\boldsymbol{T}}$. The trainable parameters $\boldsymbol{\theta} = (\boldsymbol{\alpha},\ \boldsymbol{\beta})$. $\widetilde{L}$ represents the number of tokens in $\widetilde{\boldsymbol{T}}$. $\mathcal{L}_{\mathrm{CE}}$ is the CE loss, denoted as

$$\mathcal{L}_{\mathrm{CE}}(\widetilde{\boldsymbol{T}},\ \widehat{\widetilde{\boldsymbol{T}}};\ \boldsymbol{\theta}) = -\sum_{\widetilde{l}=1}^{\widetilde{L}} p(\widetilde{t_{\widetilde{l}}}) \log p(\widehat{\widetilde{t_{\widetilde{l}}}}), \tag{6}$$

where $p(\widetilde{t_{\widetilde{l}}})$ and $p(\widehat{\widetilde{t_{\widetilde{l}}}})$ are the true and predicted probabilities of token $\widetilde{t_{\widetilde{l}}}$ and $\widehat{\widetilde{t_{\widetilde{l}}}}$, respectively. The token w_e belongs to a vocabulary group containing E tokens and $w_e \neq \widetilde{t_{\widetilde{l}}}$. Therefore, $f_w\left(w_e\right)$ describes the confidence level associated with $\widehat{\widetilde{t_{\widetilde{l}}}}$ matching w_e instead of $\widetilde{t_{\widetilde{l}}}$, which can be written as

$$f_w\left(w_e\right) = -\sum_{e=1, w_e \neq \widetilde{t_{\widetilde{l}}}}^{E} \frac{\kappa}{E-1} p(w_e) \log p(\widehat{\widetilde{t_{\widetilde{l}}}}). \tag{7}$$

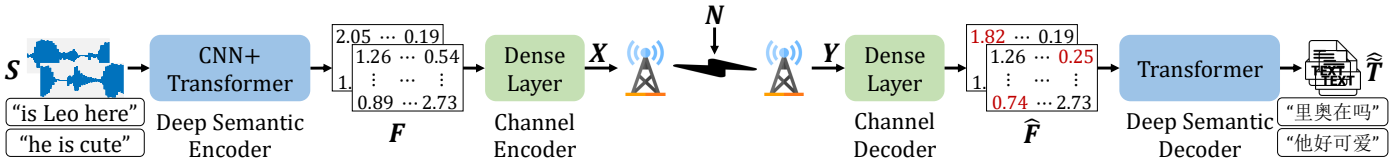
The intuition behind loss $\mathcal{L}_{\mathrm{LSR-CE}}$ is to introduce a level of confusion in predicting the target text, which enables the training uncertainty but ultimately improves the prediction accuracy in the testing stage. Algorithm 1 illustrates the algorithm for training the DeepSC-S2T.

B. Ross-S2T

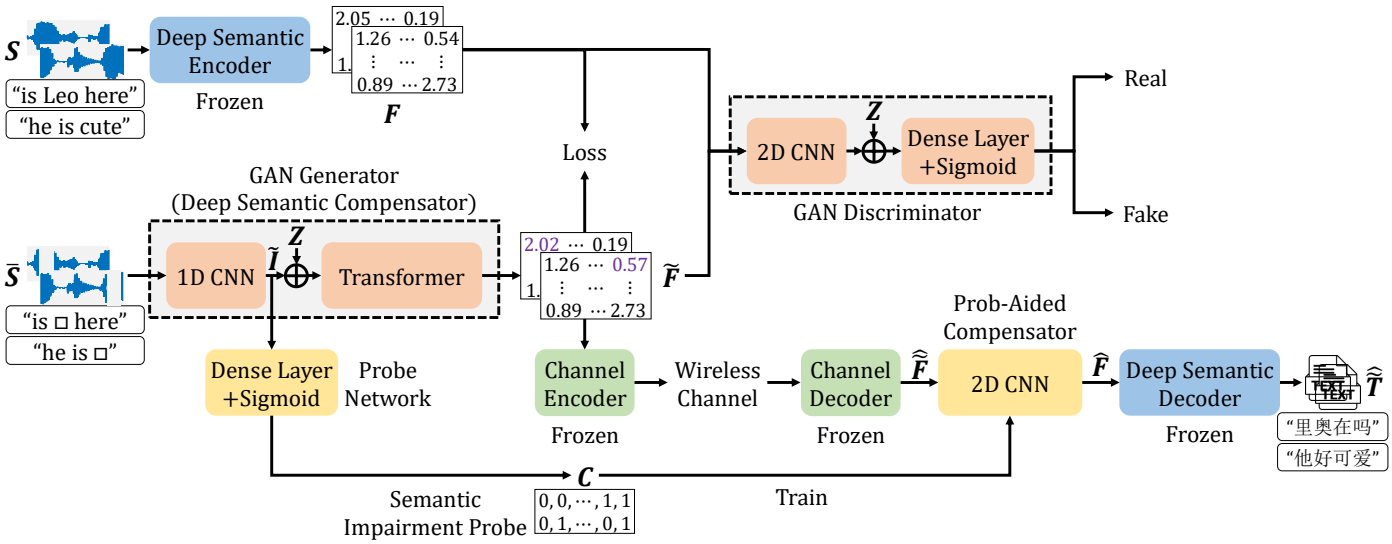
As aforementioned, the deep semantic compensator is responsible for estimating the damaged semantic information in $\overline{\boldsymbol{S}}$, extracting the deep semantic features with the least dissimilarity to $\boldsymbol{F}$, and returning the semantic impairment probe matrix to record the positional information of corrupted deep semantic features. In the second training stage, we propose a dual-compensator mechanism, including the GAN-enabled deep semantic compensator at the transmitter and the probe-aided compensator at the receiver. Particularly, the trained deep semantic encoder is leveraged to obtain $\boldsymbol{F}$ as the real data for the discriminator. The generator is developed to process the corrupted speech, $\overline{\boldsymbol{S}}$, and generate the fake data $\widetilde{\boldsymbol{F}}$ to fool the discriminator by adopting the 1D CNN module followed by the latent space, $\boldsymbol{Z}$, and the transformer module. Note that the intermediate semantic representation, $\widetilde{\boldsymbol{I}}$, are attained as the output of the 1D CNN module. The discriminator distinguishes whether the input data is real or fake, which incorporates the 2D CNN module prior to latent space $\boldsymbol{Z}$ followed by the dense layer and the sigmoid layer. Denote the trainable NN parameters of the discriminator and the generator as $\boldsymbol{\gamma}_{\mathrm{D}}$ and $\boldsymbol{\gamma}_{\mathrm{G}}$, respectively, i.e., $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{\mathrm{D}},\ \boldsymbol{\gamma}_{\mathrm{G}})$. Then, the loss function adopted for training the discrimination can be expressed as

$$\mathcal{L}_{\mathrm{D}}(\boldsymbol{S},\ \overline{\boldsymbol{S}};\ \boldsymbol{\gamma}_{\mathrm{D}}) = \frac{1}{2}\left(\mathfrak{T}_{\mathrm{D}}\left(\mathfrak{T}_{\mathrm{S}}\left(\boldsymbol{S}\right)\right) - 1\right)^2 + \frac{1}{2}\left(\mathfrak{T}_{\mathrm{D}}(\mathfrak{T}_{\mathrm{G}}(\overline{\boldsymbol{S}}))\right)^2, \tag{8}$$

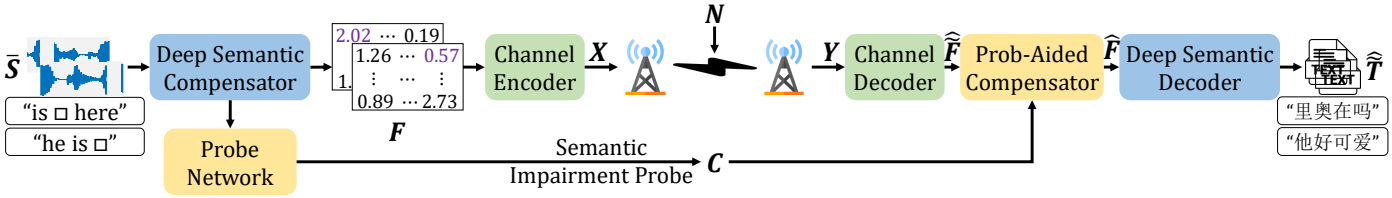
where $\mathfrak{T}_{\mathrm{D}}(\cdot)$ and $\mathfrak{T}_{\mathrm{G}}(\cdot)$ are the discriminator and the generator, respectively.



(a) First training stage: Train DeepSC-S2T with clear speech input.



(b) Second training stage: Train Ross-S2T with corrupted speech input.



(c) Testing stage: Test Ross-S2T with corrupted speech input.

Fig. 2: The model structure of proposed Ross-S2T for robust semantic communications with clear and corrupted speech inputs.

Furthermore, the γ_{G} can be updated through the loss function as follows,

$$\begin{aligned}\mathcal{L}_{\mathrm{G}}(\boldsymbol{S},\ \overline{\boldsymbol{S}};\ \gamma_{\mathrm{G}}) &= \frac{1}{2}\xi\mathcal{L}_{\mathrm{MSE}} + \frac{1}{2}\mathcal{L}_{\mathrm{ADV}} \\ &= \frac{1}{2}\xi\left(\boldsymbol{F}-\widetilde{\boldsymbol{F}}\right)^2 + \frac{1}{2}\left(\mathfrak{T}_{\mathrm{D}}(\mathfrak{T}_{\mathrm{G}}(\overline{\boldsymbol{S}}))-1\right)^2,\end{aligned} \tag{9}$$

where ξ is a hyperparameter to balance the weights of the MSE loss, $\mathcal{L}_{\mathrm{MSE}}$, and the adversarial loss, $\mathcal{L}_{\mathrm{ADV}}$.

As a result, the trained generator can create realistic data to deceive the discriminator and produce compensated deep semantic features, $\widetilde{\boldsymbol{F}}$, that closely resemble $\boldsymbol{F}$. The details for training the deep semantic compensator are presented in Algorithm 2.

Furthermore, a probe network is established at the transmission by taking $\widetilde{\boldsymbol{I}}$ as the input and providing the semantic impairment probe, $\boldsymbol{C}$, written as

$$\boldsymbol{C} = \mathfrak{T}_{\mathrm{PN}}(\widetilde{\boldsymbol{I}}) \quad \text{w.r.t.} \quad \boldsymbol{\delta}, \tag{10}$$

where $\mathfrak{T}_{\mathrm{PN}}(\cdot)$ is the probe network and $\boldsymbol{\delta}$ is NN parameters.

Algorithm 1 Training algorithm of DeepSC-S2T.

Initialization: Initialize trainable parameters $\boldsymbol{\theta}$.

1: **Input:** Clear speech input $\boldsymbol{S}$ and accurate target text $\widetilde{\boldsymbol{T}}$ from tainset $\mathfrak{S}$, fading channel $\boldsymbol{H}$, Gaussian noise $\boldsymbol{N}$.
2: **while** loss $\mathcal{L}_{\mathrm{LSR-CE}}(\boldsymbol{\theta})$ is not converged **do**
3: $\quad\mathfrak{T}_{\mathrm{C}}(\mathfrak{T}_{\mathrm{S}}(\boldsymbol{S})) \rightarrow \boldsymbol{X}$.
4: $\quad$Transmit $\boldsymbol{X}$ over $\boldsymbol{H}$ and receive $\boldsymbol{Y}$ via (2).
5: $\quad\mathfrak{T}_{\mathrm{S}}^{-1}(\mathfrak{T}_{\mathrm{C}}^{-1}(\boldsymbol{Y})) \rightarrow \widehat{\widetilde{\boldsymbol{T}}}$.
6: $\quad$Compute $\mathcal{L}_{\mathrm{LSR-CE}}(\boldsymbol{\theta})$ to update $\boldsymbol{\theta}$.
7: **end while**
8: **Output:** Trained $\mathfrak{T}_{\mathrm{S}}(\cdot)$, $\mathfrak{T}_{\mathrm{C}}(\cdot)$, $\mathfrak{T}_{\mathrm{C}}^{-1}(\cdot)$, and $\mathfrak{T}_{\mathrm{S}}^{-1}(\cdot)$.

The loss function for training the $\mathfrak{T}_{\mathrm{PN}}(\cdot)$ is denoted as

$$\mathcal{L}_{\mathrm{PN}}(\boldsymbol{I},\ \widetilde{\boldsymbol{I}},\ \boldsymbol{C};\ \boldsymbol{\delta}) = \sum_{l'=1}^{L'}\left(i_{l'} - c_{l'}\widetilde{i}_{l'}\right)^2, \tag{11}$$

where $\boldsymbol{I}$ is the intermediate semantic representation extracted from the clear speech.

Algorithm 2 Training algorithm of GAN-enabled deep semantic compensator with corrupted speech input.

Initialization: Initialize trainable parameters $\boldsymbol{\gamma}_{\mathrm{D}}$ and $\boldsymbol{\gamma}_{\mathrm{G}}$.

Input: Clear speech input $\boldsymbol{S}$ and corrupted speech input $\overline{\boldsymbol{S}}$ from tainset $\mathfrak{S}$.
Obtain the trained $\mathfrak{T}_{\mathrm{S}}(\cdot)$ from Algorithm 1.
while losses $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\gamma}_{\mathrm{D}})$ and $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\gamma}_{\mathrm{G}})$ are not converged **do**
  Generate real data via $\mathfrak{T}_{\mathrm{S}}(\boldsymbol{S}) \rightarrow \boldsymbol{F}$.
  Generate fake data via $\mathfrak{T}_{\mathrm{G}}(\overline{\boldsymbol{S}}) \rightarrow \widetilde{\boldsymbol{F}}$.
  Distinguish real data via $\mathfrak{T}_{\mathrm{D}}(\boldsymbol{F})$.
  Distinguish fake data via $\mathfrak{T}_{\mathrm{D}}(\widetilde{\boldsymbol{F}})$.
  Compute $\mathcal{L}_{\mathrm{D}}(\boldsymbol{\gamma}_{\mathrm{D}})$ and update $\boldsymbol{\gamma}_{\mathrm{D}}$.
  for each update of generator $\mathfrak{T}_{\mathrm{G}}(\cdot)$ **do**
    Generate real data via $\mathfrak{T}_{\mathrm{S}}(\boldsymbol{S}) \rightarrow \boldsymbol{F}$.
    Generate fake data via $\mathfrak{T}_{\mathrm{G}}(\overline{\boldsymbol{S}}) \rightarrow \widetilde{\boldsymbol{F}}$.
    Distinguish fake data via $\mathfrak{T}_{\mathrm{D}}(\widetilde{\boldsymbol{F}})$.
    Compute $\mathcal{L}_{\mathrm{G}}(\boldsymbol{\gamma}_{\mathrm{G}})$ to update $\boldsymbol{\gamma}_{\mathrm{G}}$.
  end for
end while
Output: Trained $\mathfrak{T}_{\mathrm{D}}(\cdot)$ and $\mathfrak{T}_{\mathrm{G}}(\cdot)$.

To further enhance the fidelity of the received deep semantic features, $\widehat{\widetilde{\boldsymbol{F}}}$, the learned semantic impairment probe, $\boldsymbol{C}$, is utilized to identify the corrupted deep semantic feature in $\widehat{\widetilde{\boldsymbol{F}}}$ and the CNN-enabled probe-aided compensator commits to reducing semantic errors between the identified corrupted features and the corresponding accurate features. The compensated deep semantic features, $\widehat{\boldsymbol{F}}$, can be expressed as

$$\widehat{\boldsymbol{F}} = \mathfrak{T}_{\mathrm{PC}}(\widehat{\widetilde{\boldsymbol{F}}},\ \boldsymbol{C}) \quad \text{w.r.t.} \quad \boldsymbol{\zeta}, \tag{12}$$

where $\mathfrak{T}_{\mathrm{PC}}(\cdot)$ indicates the probe-aided compensator and $\boldsymbol{\zeta}$ is its NN parameters.

Besides, the $\boldsymbol{\zeta}$ can be updated by

$$\mathcal{L}_{\mathrm{PC}}(\widetilde{\boldsymbol{T}},\ \widehat{\widetilde{\boldsymbol{T}}},\ \boldsymbol{C};\ \boldsymbol{\zeta}) = -\sum_{l^{\backprime}=1,\ c_{l^{\backprime}}\neq 0}^{L^{\backprime}} p(\widetilde{t}_{l^{\backprime}}) \log p(\widehat{\widetilde{t}}_{l^{\backprime}}), \tag{13}$$

where $l^{\backprime}$ indicates the position corresponding to the semantic impairment probe with the value of one.

According to the compensation operation at the receiver, the understandability of translated text, $\widehat{\widetilde{\boldsymbol{T}}}$ can be improved compared to scenarios where the received features are directly fed into the deep semantic decoder. The details for training the probe network and the probe-aided compensator are introduced in Algorithm 3.

As shown in Fig. 2 (c), the trained GAN generator and probe network are invoked to acquire features $\widetilde{\boldsymbol{F}}$ and semantic nose probe $\boldsymbol{C}$, respectively, according to the corrupted speech input. In addition, the trained channel encoder and channel decoder from the first training stage are utilized to enable robust semantic communications. The probe-aided compensator from the second training stage calibrates the corrupted deep semantic features, and the deep semantic decoder generates text in the target language.

Algorithm 3 Training algorithm of probe network and probe-aided compensator.

Initialization: Initialize trainable parameters $\boldsymbol{\delta}$ and $\boldsymbol{\zeta}$.

Input: Corrupted speech input $\overline{\boldsymbol{S}}$ from tainset $\mathfrak{S}$.
Obtain the trained $\mathfrak{T}_{\mathrm{C}}(\cdot)$, $\mathfrak{T}_{\mathrm{C}}^{-1}(\cdot)$, and $\mathfrak{T}_{\mathrm{S}}^{-1}(\cdot)$ from Algorithm 1, the trained $\mathfrak{T}_{\mathrm{G}}(\cdot)$ from Algorithm 2.
while loss $\mathcal{L}_{\mathrm{PN}}(\boldsymbol{\delta})$ and loss $\mathcal{L}_{\mathrm{PC}}(\boldsymbol{\zeta})$ are not converged **do**
  Attain intermedia semantic representation $\widetilde{\boldsymbol{I}}$.
  Obtain semantic impairment probe via $\mathfrak{T}_{\mathrm{PN}}(\widetilde{\boldsymbol{I}}) \rightarrow \boldsymbol{C}$.
  Generate deep semantic features via $\mathfrak{T}_{\mathrm{G}}(\overline{\boldsymbol{S}}) \rightarrow \widetilde{\boldsymbol{F}}$.
  $\mathfrak{T}_{\mathrm{C}}(\widetilde{\boldsymbol{F}}) \rightarrow \boldsymbol{X}$.
  Transmit $\boldsymbol{X}$ over $\boldsymbol{H}$ and receive $\boldsymbol{Y}$ via (2).
  $\mathfrak{T}_{\mathrm{C}}^{-1}(\boldsymbol{Y}) \rightarrow \widehat{\widetilde{\boldsymbol{F}}}$.
  Compensate $\widehat{\widetilde{\boldsymbol{F}}}$ via $\mathfrak{T}_{\mathrm{PC}}(\widehat{\widetilde{\boldsymbol{F}}},\ \boldsymbol{C}) \rightarrow \widehat{\boldsymbol{F}}$.
  Obtained target text via $\mathfrak{T}_{\mathrm{S}}^{-1}(\widehat{\boldsymbol{F}}) \rightarrow \widehat{\widetilde{\boldsymbol{T}}}$.
  Compute $\mathcal{L}_{\mathrm{PN}}(\boldsymbol{\delta})$ and $\mathcal{L}_{\mathrm{PC}}(\boldsymbol{\zeta})$ to update $\boldsymbol{\delta}$ and $\boldsymbol{\zeta}$.
end while
Output: Trained $\mathfrak{T}_{\mathrm{PN}}(\cdot)$ and $\mathfrak{T}_{\mathrm{PC}}(\cdot)$.

IV. Numerical Results

In the experiments, the corpus *CoVoST 2* is used as the clear speech dataset. To create the corrupted speech dataset, the clear speech is engulfed by semantic impairments. The bilingual evaluation understudy (BLEU) and semantic textual similarity (STS) [14] are employed as the efficient performance metrics to evaluate the performance of the proposed Ross-S2T over AWGN channels and Rayleigh fading channels with the accurate channel state information (CSI). Moreover, we chose English as the source language and Chinese as the target language.

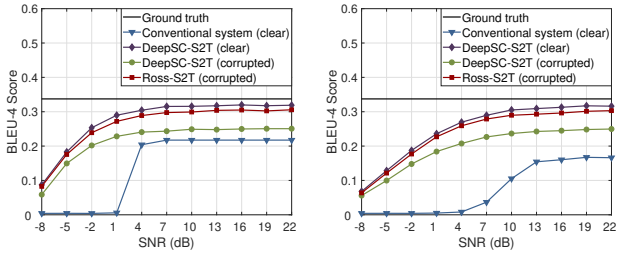
A. Simulation Settings

The DeepSC-S2T and the GAN-enabled deep semantic compensator are trained separately to cope with the dynamic speech input. In the DeepSC-S2T, the deep semantic encoder consists of five CNN modules and eight transformer modules, and the channel encoder/decoder has three dense layers with 1024 units. Six transformer modules are utilized in the deep semantic decoder. Moreover, the generator of the Ross-S2T is constructed by seven CNN modules and two dense layers followed by eight transformer modules, and the discriminator includes six transformer modules and three CNN modules followed by a dense layer and a softmax layer. The hyper-parameters $\kappa = 0.95$ and $\xi = 10$. The batch size is 128 and the number of warm-up steps is 4000. The details of various NN settings are summarized in Table I.

The BLEU results of the Ross-S2T are shown in Fig. 3, where the ground truth results are obtained by feeding the clear speech into an S2TT pipeline constructed from the conformer and the BART, and a benchmark is provided by a conventional speech transmission system consisting of the adaptive multi-rate code, the polar code, and the 64-QAM. From the figure, the DeepSC-S2T dramatically outperforms

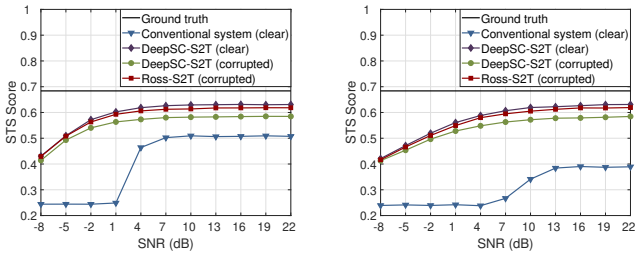
TABLE I: Parameter settings of the proposed Ross-S2T.

			Layer Name	Parameters	Activation
Ross-S2T	**DeepSC-S2T**	**Deep Semantic Encoder**	7×1D CNN modules	512 channels	None
			6×Transformer Modules	128 (8 heads)	GELU
		Channel Encoder	2×Dense Layers	1024 units	ReLU
		Channel Decoder	2×Dense Layers	1024 units	ReLU
		Deep Semantic Decoder	6×Transformer Modules	128 (8 heads)	GELU
	GAN	**Generator**	7×1D CNN modules	512 channels	None
			8×Transformer Modules	128 (8 heads)	GELU
		Discriminator	5×2D CNN modules	512 channels	None
			1×Dense Layer	1024 units	Sigmoid
	Probe Network		3×Dense Layer	1024 units	Sigmoid
	Probe-Aided Compensator		5×2D CNN modules	512 channels	None



(a) AWGN channels (b) Rayleigh channels

Fig. 3: Simulation results of BLEU scores (4-grams).



(a) AWGN channels (b) Rayleigh channels

Fig. 4: Simulation results of STS scores.

the conventional approach in the context of clear speech. However, the BLEU scores obtained by the DeepSC-S2T are lower than the proposed Ross-S2T under the observed channel environments when contending with the corrupted speech, verifying the superiority of the dual-compensator mechanism.

Fig. 4 presents the STS comparison of various methods. From the figure, the Ross-S2T attains the STS score of around 0.6 under the AWGN channels when SNR=1 dB, while the STS score of the conventional system falls below 0.3. Moreover, the DeepSC-S2T manifests a significantly inferior capability in recovering the impaired semantic information within the corrupted speech compared to the Ross-S2T.

V. Conclusions

In this paper, we study the robust semantic communications for speech transmission, named Ross-S2T, to support end-to-end speech-to-text translation (S2TT). Particularly, a deep semantic encoder is developed to learn textual semantic features related to another language from the clear speech, which enables the deep semantic exchange to achieve S2TT at the receiver. Moreover, a generative adversarial network (GAN)-enabled deep semantic compensator and a probe-aided compensator are tailored for corrupt speech scenarios by estimating the impaired semantic information and attaining as accurate deep semantic features as possible. Simulation results demonstrated the superiority of the proposed Ross-S2T to serve S2TT tasks and suppress semantic impairments.

References

[1] W. Tong and G. Y. Li, "Nine challenges in artificial intelligence and wireless communications for 6G," *IEEE Wireless Commun.*, vol. 29, no. 4, pp. 140–145, May 2022.

[2] Z. Qin, X. Tao, J. Lu, W. Tong, and G. Y. Li, "Semantic communications: Principles and challenges," *arXiv preprint arXiv:2201.01389*, Dec. 2021.

[3] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Champaign, IL, USA: Univ. Illinois Press, 1949.

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

[5] 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 *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Rio de Janeiro, Brazil, Dec. 2022, pp. 2704–2709.

[6] Z. Weng and Z. Qin, "Semantic communication systems for speech transmission," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp. 2434–2444, Jun. 2021.

[7] Z. Xiao, S. Yao, J. Dai, S. Wang, K. Niu, and P. Zhang, "Wireless deep speech semantic transmission," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Rhodes Island, Greece, May 2023.

[8] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 55–71, Nov. 2023.

[9] P. Jiang, C.-K. Wen, S. Jin, and G. Y. Li, "Wireless semantic communications for video conferencing," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 230–244, Nov. 2022.

[10] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, Jul. 2022.

[11] Z. Weng, Z. Qin, X. Tao, C. Pan, G. Liu, and G. Y. Li, "Deep learning enabled semantic communications with speech recognition and synthesis," *IEEE Trans. Wireless Commun.*, vol. 22, no. 9, pp. 6227–6240, Feb. 2023.

[12] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Trans. Commun.*, vol. 70, no. 8, pp. 5181–5192, Jun. 2022.

[13] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," *IEEE Trans. Commun.*, pp. 1–1, Feb. 2024.

[14] E. Agirre, C. Banea, C. Cardie, D. Cer, M. Diab, A. Gonzalez-Agirre, W. Guo, R. Mihalcea, G. Rigau, and J. Wiebe, "SemEval-2014 task 10: Multilingual semantic textual similarity," in *Proc. Int. Workshop Semantic Eval. (SemEval)*, Dublin, Ireland, Aug. 2014, pp. 81–91.