Semantic Communication Systems for Speech Transmission

Zhenzi Weng, *Student Member, IEEE,* and Zhijin Qin, *Member, IEEE*

***Abstract*—Semantic communications could improve the transmission efficiency significantly by exploring the input semantic information. Motivated by the breakthroughs in deep learning (DL), we make an effort to recover the transmitted speech signals in the semantic communication systems, which minimizes the error at the semantic level rather than the bit level or symbol level as in the traditional communication systems. Particularly, we design a DL-enabled semantic communication system for speech signals, named DeepSC-S. Based on an attention mechanism employing squeeze-and-excitation (SE) networks, DeepSC-S is able to identify the essential speech information and assign high values to the weights corresponding to the essential information when training the neural network. Moreover, in order to facilitate the proposed DeepSC-S to cater to dynamic channel environments, we dedicate to find a general model to cope with various channel conditions without retraining. Furthermore, to verify the model adaptation in practice, we investigate DeepSC-S in the telephone systems as well as the multimedia transmission systems, which usually requires higher data rates and lower transmission latency. The simulation results demonstrate that our proposed DeepSC-S achieves higher system performance than the traditional communications in both telephone systems and multimedia transmission systems by comparing the speech signals metrics, signal-to-distortion ration and perceptual evaluation of speech distortion. Besides, DeepSC-S is more robust to channel variations than the traditional approaches, especially in the low signal-to-noise (SNR) regime.**

***Index Terms*—Deep learning, semantic communication, speech transmission, squeeze-and-excitation networks.**

I. Introduction

INTELLIGENT communications have attracted intensive attention for traditional communication systems [1]. Inspired by the success in various areas, deep learning (DL) has been considered as a promising candidate for communications to achieve higher system performance with more intelligence [2]. Particularly, DL has shown its great potentials to solve the existing technical problems in both physical layer communications [3]–[5] and wireless resource allocations [6], [7].

Typically, a DL-based communication system is designed to reduce the complexity and/or improve the system performance, by merging one or multiple communication modules in the traditional block-wise architecture and using deep neural networks (DNN) with trainable parameters to represent the intelligent transceiver. However, even if the communication systems utilizing DL technique yield better performance and/or lower complexity for some scenarios and conditions, most of the literature focus on the performance improvement at the bit or symbol level, which usually takes bit-error rate (BER) or symbol-error rate (SER) as the performance metrics. Particularly, the major task in the traditional communication systems and the developed DL-enabled systems, is to recover the transmitted message accurately and effectively, represented by digital bit sequences. In the past decades, such type of wireless communication systems have experienced significant development from the first generation (1G) to the fifth generation (5G) and the system capacity is approaching Shannon limit.

Shannon and Weaver [8] categorized communications into three levels:

- *Level A*: how accurately can the symbols of communication be transmitted? (The technical problem.)
- *Level B*: how precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)
- *Level C*: how effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)

This indicates the feasibility to transmit the semantic information, instead of the bits or symbols, to achieve higher system efficiency. Besides, due to the increasing deployment of intelligent IoT applications, e.g., human-computer interactions and machine-machine communications, semantic-irrelative communications are no longer ideal in the future. Motivated by this, researchers have dedicated to develop a new system to process and exchange semantic information for more efficient communications.

Semantic theory, in contrast to information theory exploited in existing communication systems, takes into account the meaning and veracity of source information because they can be both informative and factual [9], which facilitates semantic communication systems to recover information at the receiver via minimizing the meaning difference between the input and the recovered signals instead of BER or SER. However, the exploration of semantic communications has gone through decades of stagnation since it was first identified because some fundamental problems, i.e., lack of mathematical model, cannot be formulated and solved properly when semantic information exchange is considered [10], e.g., *how to define the efficiency and reliability in semantic communication?* According to the recent efforts in [11], semantic data can be compressed to proper size for transmission using a lossless method by utilizing the semantic relationship between different messages, while the traditional lossless source coding is to represent a signal with the minimum number of binary bits by exploring the dependencies or statistical properties of input signals. In addition, the end-to-end (E2E) communication

Zhenzi Weng and Zhijin Qin are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK (email: zhenzi.weng@qmul.ac.uk, z.qin@qmul.ac.uk).

systems has been developed in [12] in order to address the bottlenecks in traditional block-wise communication systems, that is sub-optimal, since the conventional signal processing is hard to capture many imperfections and non-linearities in the practical channel environment. Inspired by this, different types of sources have been considered in recent investigation on E2E semantic communication systems, which mainly focus on the image and text transmission [23]–[31]. The investigation on semantic communication for speech signals transmission is still missed.

In the area of speech signal processing, the cutting edge DL applications are developed to convert speech signals into text information, e.g., automatic speech recognition (ASR). The core of ASR is to generate corresponding texts by interacting speech signals with a acoustic model by mapping each phoneme into a single alphabet, and then concatenating all alphabets into a understandable word sequence via a language model, which pays no attention to the characteristics of speech signals, e.g., the speaking speed and tone [13]. However, our work is to recover speech signals, which includes the recovery of speech characteristics. Thus, the language model based ASR technologies are not applicable as the speech characteristics are abandoned when speech signals are converted into texts and the process is irreversible. Moreover, most DL algorithms pre-process speech signals to obtain magnitude, spectra, or Mel-Frequency Cepstrum by various operations, such as discrete cosine transform (DCT), before feeding into a learning system. Such operations are employed to capture the unique features of speech signals, e.g., inconsistent speaking speeds of a person speaking in different moments are inconsistent, different frequencies of female and male voices, and distinct tones of persons at different ages, which runs counter to the motivation of intelligence. Therefore, a DL algorithm for feature learning directly from speech signal is desired. The DL based semantic communication systems by learning semantic information of speech signals directly are of great interest and importance for the next generation communication systems.

In this paper, we explore the semantic systems for speech signals by utilizing DL technique. Particularly, a DL-enabled semantic communication system for speech signals, named DeepSC-S, is proposed by learning and extracting speech signals, and then recovering them at the receiver from the received features directly. The main contributions of this article can be summarized as fourfold:

- A novel semantic communication system for speech signals, named DeepSC-S, is first proposed, which treats the transmitter and the receiver as two trainable DNNs, and jointly designs the speech coding and the channel coding to deal with source distortion and channel effects.
- Particularly, in the proposed DeepSC-S, the squeeze-and-excitation (SE) networks [14] is employed to learn and extract essential speech semantic information, as well assign high values to the weights corresponding to the essential information during the training phase. By exploiting the attention mechanism based on SE networks, DeepSC-S improves the accuracy of signal recovering.
- Moreover, by training DeepSC-S under a fixed fading channel and SNR, then facilitating the trained model with good performance under testing channel conditions, the proposed DeepSC-S is highly robust to dynamic channel environments without network tuning and retraining.
- To verify the model adaptation to practical communication scenarios, the proposed DeepSC-S is applied to telephone systems and multimedia transmission systems, respectively. The performance is also verified with traditional approaches to prove its superiority. Simulation results show that DeepSC-S outperforms the traditional systems, especially in the low SNR regime.

The rest of this article is structured as follows. The related work is presented in Section II. Section III introduces the model of semantic communication system for speech transmission and the related performance metrics. In Section IV, the details of the proposed DeepSC-S is presented. Simulation results are discussed in Section V. Section VI draws conclusions.

Notation: Single boldface letters are used to represent vectors or matrices and single plain capital letters denotes integers. Given a vector $\boldsymbol{x}$, x_i indicates its i-th component, $\|\boldsymbol{x}\|$ denotes its Euclidean norm. Given a matrix $\boldsymbol{Y}$, $\boldsymbol{Y} \in \Re^{M \times N}$ indicates $\boldsymbol{Y}$ is a matrix with real values and the size is $M \times N$. Superscript swash letters refers a block in the system, e.g., $\mathcal{T}$ in $\boldsymbol{\theta}^{\mathcal{T}}$ represents the parameter of the transmitter. $\mathcal{CN}(\boldsymbol{m},\ \boldsymbol{V})$ are multivariate circular complex Gaussian distribution with mean vector $\boldsymbol{m}$ and co-variance matrix $\boldsymbol{V}$, respectively. Moreover, $\boldsymbol{a} * \boldsymbol{b}$ represents the convolution operation on the vector $\boldsymbol{a}$ and the vector $\boldsymbol{b}$.

II. Related Work

As aforementioned that E2E learning of communication systems has been developed to address the challenges in traditional communication systems, however, it analyses and improves the system performance at the bit or symbol level. Moreover, the existing applications on DL-enabled semantic communication systems are mainly based on text and image source information.

A. End-to-End Learning in Communication Systems

The DL-based E2E communication systems have achieved extremely competitive block-error rate (BLER) performance compared to the traditional baselines in various scenarios [12], e.g., uncoded binary phase shift keying (BPSK) and Hamming coded BPSK. In addition, it has shown great potentials in processing complicated communication tasks. For example, the E2E learning systems have been employed in orthogonal frequency division multiplexing (OFDM) systems [15], [16], as well as in multiple-input multiple-output (MIMO) systems [17], [18]. Besides, channel estimation is a challenging problem in the DL enabled E2E systems. In [19], reinforcement learning (RL) has been unitized to estimate channel state information (CSI) through treating the channel layer and the receiver as the *environment*, the transmitter as the *agent* to take *actions* to interact with the *environment* based on a *policy*, which has been assumed as the most cutting-edge approach while an additional reliable channel is still required to send losses back from the receiver to the transmitter during the training phase. Another novel channel agnostic solution has been proposed in [20], which replaces the realistic channels with a neural network (NN) by exploiting a conditional generative adversarial network (GAN).

Furthermore, due to the complexity of NN training in the E2E learning models, high training efficiency with low energy consumption system is desirable when employing it into the practical scenarios. Transfer learning has been considered as a promising technology for adapting E2E communication systems to cope with the uncontrollable and unpredictable channel environments by training them over a statistical channel model [21]. In addition, another appealing solution is to obtain a trained model yielding expected performance via small number of stochastic gradient descent (SGD) iterations. Particularly, a model agnostic meta-learning enabled E2E communication system has been investigated in [22], which finds a common initialization parameter achieving fast convergence after one or several iterations for various channel conditions.

B. Semantic Communications

An initial research on semantic communication systems for text information has been developed [23], which mitigates the semantic error to achieve Nash equilibrium by integrating the semantic inference and the physical layer communication to optimize the whole transceiver. However, such a text-based semantic communication system only measures the difference between the transmitted sentences and the received sentences at the word level instead of the sentence level. Thus, a further investigation about semantic communications for text transmission, named DeepSC, has been carried out in [24] to deal with the semantic error at the sentence level with various length. Powered by the Transformer [25], the semantic encoder and the channel encoder are jointly designed as a trainable autoencoder to minimize the semantic error, rather than the BER or SER as in traditional communications and to improve the system capacity. By doing so, the semantic communications for text could be realized. Moreover, the increasing deployment of smart IoT devices has required the IoT devices to implement more complicated tasks, such as training a DNN independently, which runs counter to the limited computing capability of IoT devices. Inspired by this, a lite distributed semantic communication system for text transmission, named L-DeepSC, has been proposed in [26] to address the challenge of IoT to perform the intelligent tasks by pruning parameters to reduce the size of the trained models as well as to reduce the communication cost between IoT devices and the server.

In the area of semantic communications for image information, a DL-enabled semantic communication system for image transmission, named JSCC, has been developed [27], which employs a convolutional neural network (CNN) with five convolutional layers to jointly design the source-channel encoder, and a CNN with five transposition convolutional layers at the receiver to realize the source-channel decoder. Based on JSCC, an image transmission system, integrating noiseless or noisy channel output feedback, has been investigated to improve image reconstruction [28], where the channel output backpropagates to the transmitter based on a unit delay and additive white Gaussian noise (AWGN) is added to generate a weight vector to the NN at the transmitter. By utilizing the feedback mechanism, the quality of reconstructed image is improved compare to the model without feedback signals and the traditional approaches. Similar to text transmission, IoT applications for image transmission have been carried out. Particularly, a joint image transmission-recognition system has been developed [29] by applying two DNNs as the transmitter at the IoT devices and the receiver at the server edge, which has shown the superior recognition accuracy than the traditional approaches and has great advantage of low computation resource via transfer learning. In [30], a deep joint source-channel coding architecture, name DeepJSCC, combines with network pruning technique to perform image classification at the edge sever, which facilitates the IoT services to process image with low computation complexity and reduce the requirement of transmission bandwidth. Moreover, an application based on JSCC to retrieve image at wireless edge has been proposed in [31], which aims to address transmission delay of IoT devices to send the whole quilted image by employing a DNN to realize retrieval-oriented image compression.

Given the intensive investigations of semantic communication for text and image information as well as the challenges of traditional communications for speech transmission, e.g., poor telephone communication quality at the airport, it is significant to carry out the research on speech communication systems by utilizing the semantic information.

III. System Model

In this section, we first introduce the considered system model. Besides, the details of the system model and the performance metrics are presented.

A. System Settings

The considered system will transmit the original speech signals via a NN-based speech semantic communication system, which comprises two major tasks as shown in Fig. 1: i) semantic information learning and extracting of speech signals; ii) and mitigating the effects of wireless channels. Due to the variation of speech characteristics, it is a quite challenging problem. For a practical communication scenario, the signal passing through the physical channel suffers from distortion and attenuation. Therefore, the considered DL enabled system targets to recover the original speech signals and achieve better performance than the traditional approaches while coping with complicated channel distortions.

B. Transmitter

The proposed system model is shown in Fig. 1. From the figure, the input of the transmitter is a speech sample sequence, $\boldsymbol{s} = [s_1,\ s_2,\ ...,\ s_W]$ with W samples, where s_w is w-th item in $\boldsymbol{s}$ and it is a scalar value, i.e., a positive number, a negative number, or zero. At the transmitter, the input, $\boldsymbol{s}$, is mapped into symbols, $\boldsymbol{x}$, to be transmitted over physical channels. As

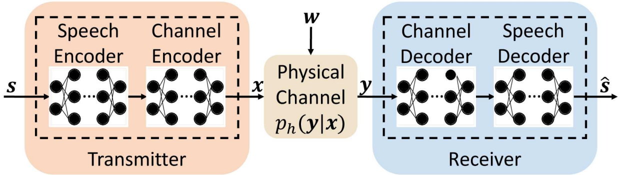


Fig. 1: The model structure of DL enabled speech semantic communication system.

shown in Fig. 1, the transmitter consists of two individual components: the *speech encoder* and the *channel encoder*, in which each component is implemented by an independent NN. Denote the NN parameters of the *speech encoder* and the *channel encoder* as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Then the encoded symbol sequence, $\boldsymbol{x}$, can be expressed as

$$\boldsymbol{x} = \mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\boldsymbol{s})), \tag{1}$$

where $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\cdot)$ and $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\cdot)$ indicate the *speech encoder* and the *channel encoder* with respect to (w.r.t.) parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Here we denote the NN parameters of the transmitter as $\boldsymbol{\theta}^{\mathcal{T}} = (\boldsymbol{\alpha},\ \boldsymbol{\beta})$.

The mapped symbols, $\boldsymbol{x}$, are transmitted over a physical channel. Note that the normalization on transmitted symbols $\boldsymbol{x}$ is required to ensure the total transmission power constraint $\mathbb{E}\left\|\boldsymbol{x}\right\|^2 = 1$.

The whole transceiver in Fig. 1 is designed for a single communication link, in which the channel layer, represented by $p_h\left(\boldsymbol{y}|\,\boldsymbol{x}\right)$, takes $\boldsymbol{x}$ as the input and produces the output as received signal $\boldsymbol{y}$. Denote the coefficients of a linear channel as $\boldsymbol{h}$, then the transmission process from the transmitter to the receiver can be modeled as

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

where $\boldsymbol{w} \sim \mathcal{CN}(0,\ \sigma^2\mathbf{I})$ indicates independent and identically distributed (i.i.d.) Gaussian noise, σ^2 is noise variance for each channel and $\mathbf{I}$ is the identity matrix.

C. Receiver

Similar to the transmitter, the receiver also consists of two cascaded parts, including the *channel decoder* and the *speech decoder*. The *channel decoder* is to mitigate the channel distortion and attenuation, and the *speech decoder* recovers speech signals based on the learned and extracted speech semantic features. Denote the NN parameters of the *channel decoder* and the *speech decoder* as $\boldsymbol{\chi}$ and $\boldsymbol{\delta}$, respectively. As depicted in Fig. 1, the decoded signal, $\widehat{\boldsymbol{s}}$, can be obtained from the received signal, $\boldsymbol{y}$, by the following operation:

$$\widehat{\boldsymbol{s}} = \mathbf{R}_{\boldsymbol{\delta}}^{\mathcal{S}}(\mathbf{R}_{\boldsymbol{\chi}}^{\mathcal{C}}(\boldsymbol{y})), \tag{3}$$

where $\mathbf{R}_{\boldsymbol{\chi}}^{\mathcal{C}}(\cdot)$ and $\mathbf{R}_{\boldsymbol{\delta}}^{\mathcal{S}}(\cdot)$ indicate the *channel decoder* and the *speech decoder* w.r.t. parameters $\boldsymbol{\chi}$ and $\boldsymbol{\delta}$, respectively. Denote the NN parameter of the receiver as $\boldsymbol{\theta}^{\mathcal{R}} = (\boldsymbol{\chi},\ \boldsymbol{\delta})$.

The objective of the whole transceiver system is to recover speech signals as close as to the original, which causes two challenges. The first one is the design of efficient and intelligent *speech encoder/decoder* by utilizing the semantic information to recover speech signals, especially under the poor channel conditions, such as the low SNR regime. The second one is the design of the *channel encoder/decoder* to alleviate symbol errors caused by the physical channels via adding redundancy information. For the traditional communications, the advanced channel coding techniques are achieved at the bit level to target a low BER/SER. However, the bit-to-symbol transformation is not involved in our proposed system. The raw speech signals are directly mapped into a transmitted symbol stream by the *speech encoder* and the *channel encoder*, and recover it at the receiver via inverse operations. Thus, we treat the speech recovery process as a signal reconstruction task to minimize the errors between the signal values in $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$ by exploiting the characteristics of speech signals, then mean-squared error (MSE) is used as the loss function in our system to measure the difference between $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$, denoted as

$$\mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}},\ \boldsymbol{\theta}^{\mathcal{R}}) = \frac{1}{W}\sum_{w=1}^{W}\left(s_w - \widehat{s}_w\right)^2, \tag{4}$$

where s_w and $\widehat{s}_w$ indicate the w-th element of the vectors $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$, respectively. W is the length of these two vectors.

Assume that the NN models of the whole transceiver are differentiable w.r.t. the corresponding parameters, which can be optimized via gradient descent based on the MSE loss function. It is worth to mention that the *speech encoder/decoder* and the *channel encoder/decoder* are jointly designed. Besides, given prior CSI, both parameters sets $\boldsymbol{\theta}^{\mathcal{T}}$ and $\boldsymbol{\theta}^{\mathcal{R}}$ can be adjusted at the same time. Denote the NN parameter of the whole system model as $\boldsymbol{\theta}$, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\mathcal{T}},\ \boldsymbol{\theta}^{\mathcal{R}})$, we adopt the SGD algorithm to train task in this paper. Then iteratively updates on the parameters $\boldsymbol{\theta}$ follows

$$\boldsymbol{\theta}^{(i+1)} \leftarrow \boldsymbol{\theta}^{(i)} - \eta\nabla_{\boldsymbol{\theta}^{(i)}}\mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}},\ \boldsymbol{\theta}^{\mathcal{R}}), \tag{5}$$

where $\eta > 0$ is a learning rate and ∇ indicates the differential operator.

D. Performance Metrics

In our model, the system is committed to reconstruct the raw speech signals. Hence, the signal-to-distortion ration (SDR) [32] is employed to measure the $\mathcal{L}_2$ error between $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$, which is one of the commonly used metric for speech transmission and can be expressed as

$$SDR = 10\log_{10}\left(\frac{\left\|\boldsymbol{s}\right\|^2}{\left\|\boldsymbol{s} - \widehat{\boldsymbol{s}}\right\|^2}\right). \tag{6}$$

The higher SDR represents the speech information is recovered with better quality, i.e., easier to understand for human beings. According to (4), MSE loss could reflect the goodness of SDR. The lower the MSE, the higher the SDR.

Perceptual evaluation of speech distortion (PESQ) [33] is another metric for the quality of speech signals at the receiver, which takes the short memory in human perception into consideration. PESQ is a speech quality assessment model combing the perceptual speech quality measure (PSQM) and perceptual analysis measurement system (PAMS), which is adopted in International Telecommunication Union (ITU-T) recommendation P.862 [34]. PESQ is a good candidate for evaluating the quality of speech messages under various conditions, e.g., background noise, analog filtering, and variable delay, by scoring the speech quality range from -0.5 to 4.5.

IV. Proposed Semantic Communication System for Speech Signals

To address the aforementioned challenges, we design a DL-enabled speech semantic communication system, named DeepSC-S. Specifically, an attention-based two-dimension (2D) CNN is used for the *speech encoder/decoder* and a 2D CNN is adopted for the *channel encoder/decoder*. The details of the developed DeepSC-S will be introduced in this section.

A. Model Description

As shown in Fig. 2, the input of the proposed DeepSC-S, denoted as $\boldsymbol{S} \in \Re^{B\times W}$, is the set of speech sample sequences, $\boldsymbol{s}$, which are drawn from the speech dataset, $\mathfrak{S}$, and B is the batch size. $\mathfrak{S}$ consists of considerable speech signals, which is collected by recording the speakings from different persons. The input sample sequences set, $\boldsymbol{S}$, are framed into $\boldsymbol{m} \in \Re^{B\times F\times L}$ for training before passing through an attention-based encoder, i.e., the *speech encoder*, where F indicates the number of frames and L is the length of each frame. Note that the framing operation only reshapes $\boldsymbol{S}$ without any feature learning and extracting. The *speech encoder* directly learns the speech semantic information from $\boldsymbol{m}$ and outputs the learned features $\boldsymbol{b} \in \Re^{B\times F\times L\times D}$. The details of the *speech encoder* is detailed in part B of this section. Afterwards, the *channel encoder*, denoted as a CNN layer with 2D CNN modules, converts $\boldsymbol{b}$ into $\boldsymbol{U} \in \Re^{B\times F\times 2N}$. In order to transmit $\boldsymbol{U}$ into a physical channel, it is reshaped into symbol sequences, $\boldsymbol{X} \in \Re^{B\times FN\times 2}$, via a reshape layer.

The channel layer takes the reshaped symbol sequences, $\boldsymbol{X}$, as the input and produces $\boldsymbol{Y}$ at the receiver, which is given by

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X} + \boldsymbol{W}, \tag{7}$$

where $\boldsymbol{H}$ consists of B number of channel coefficient vectors, $\boldsymbol{h}$, and $\boldsymbol{W}$ is Gaussian noise, which includes B number of noise vectors, $\boldsymbol{w}$.

The received symbol sequences, $\boldsymbol{Y}$, is reshaped into $\boldsymbol{V} \in \Re^{B\times F\times 2N}$ before feeding into the *channel decoder*, represented by a CNN layer with 2D CNN modules. The output of the *channel decoder* is $\widehat{\boldsymbol{b}} \in \Re^{B\times F\times L\times D}$. Afterwards, an attention-based decoder, i.e., the *speech decoder*, converts $\widehat{\boldsymbol{b}}$ into $\widehat{\boldsymbol{m}} \in \Re^{B\times F\times L}$ and $\widehat{\boldsymbol{m}}$ is recovered into $\widehat{\boldsymbol{S}}$ via the inverse operation of framing, named deframing, where the size of $\widehat{\boldsymbol{S}}$ is same as that of $\boldsymbol{S}$ at the transmitter. The loss is calculated at the end of the receiver and backpropagated to the transmitter, thus, the trainable parameters in the whole system can be updated simultaneously.

B. Speech Encoder and Decoder

The core of the proposed DeepSC-S is the NN-enabled *speech encoder* and *speech decoder* based on an attention mechanism, named SE-ResNet, which is capable of learning and extracting essential information. In this work, we focus on the essential speech semantic information, e.g., the signal magnitude, which increases sharply when emphasising important message and the signal frequency, which decreases abruptly when speaking speed becomes slow to express message more clearly. Particularly, SE-ResNet is employed to identify the essential information and the weights corresponding to the essential information are assigned to high values when weight updating and adjusting during the training phase.

As shown in Fig. 3, for the SE-ResNet, a *Split* layer splits the input, $\boldsymbol{m}$, into multiples blocks, which is achieved by multiple convolution kernels, and all the blocks are concatenated. Then a *transition* layer is utilized to reduce the dimension of the concatenated blocks and the output is denoted as $\boldsymbol{p} \in \Re^{M\times N\times C}$, which consists of C features and each feature is in size of $M \times N$. For the SE layer, a *squeeze* operation is employed to aggregate the 2D spatial dimension of each input feature, then an operation, named *excitation*, intents to output the attention factor of each feature by learning the inter-dependencies of features $\boldsymbol{p}$. The output of the SE layer, $\boldsymbol{z} \in \Re^{1\times 1\times C}$, includes C number of scale coefficients, which is considered as the attention factor to scale the importance of the extracted features in $\boldsymbol{p}$ by multiplying the features corresponding essential information in $\boldsymbol{p}$ with the high scale coefficients in $\boldsymbol{z}$. By doing so, the weights of $\boldsymbol{m}$ are reassigned, i.e., the weights corresponding to essential speech information are paid more attention. Note that the SE layer is considered as an independent unit and one or multiple SE-ResNet modules can be sequentially connected. With more SE-ResNet modules, the performance of feature learning and extracting to essential information will improve, however, it also increases computational complexity. Therefore, a tradeoff between the learning performance and complexity should be considered during the training. Additionally, residual network is adopted to alleviate the problem of gradient vanishing due to the network depth by adding $\boldsymbol{m}$ into the output of the SE-ResNet module, as shown in Fig. 3.

Particularly, the *speech encoder* is comprised by multiple SE-ResNet modules to convert input $\boldsymbol{m}$ into $\boldsymbol{b}$, corresponding to Fig. 2. For the *speech decoder*, in addition to several SE-ResNet modules, the *last layer*, including a 2D CNN module with one single kernel, is utilized to reduce the output size of the *speech decoder*, $\widehat{\boldsymbol{m}}$, as the sizes of $\boldsymbol{m}$ and $\widehat{\boldsymbol{m}}$ should be equal.

C. Model Training and Testing

Based on the prior knowledge of CSI, the transmitter and receiver parameters, $\boldsymbol{\theta}^{\mathcal{T}}$ and $\boldsymbol{\theta}^{\mathcal{R}}$, can be updated simultaneously. As aforementioned, the objective of the proposed DeepSC-S is to train a model to capture the essential information in speech signals and make it to work well under various channels and a wide SNR regime.

1) Training Stage: As in Fig. 2, the training algorithm of DeepSC-S is described in Algorithm 1. During the training stage, in order to facilitate the fast MSE loss convergence, the NN parameters, $\boldsymbol{\theta} = (\boldsymbol{\theta}^{\mathcal{T}},\ \boldsymbol{\theta}^{\mathcal{R}})$, are initialized by a variance scaling initializer, instead of 0. Besides, for achieving a valid

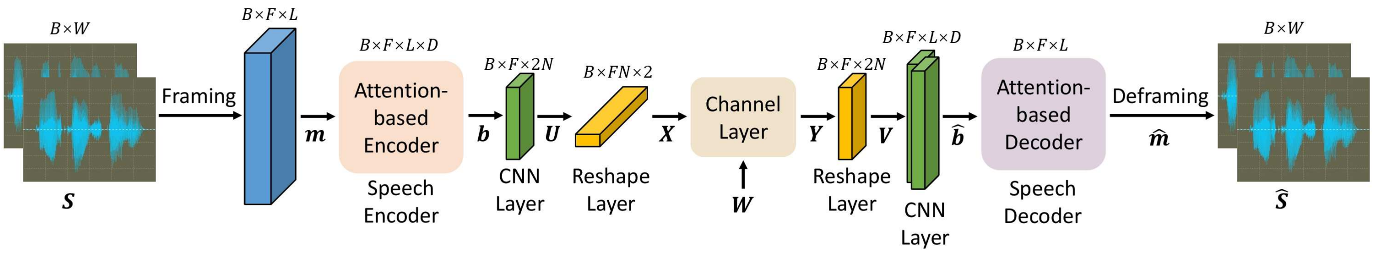
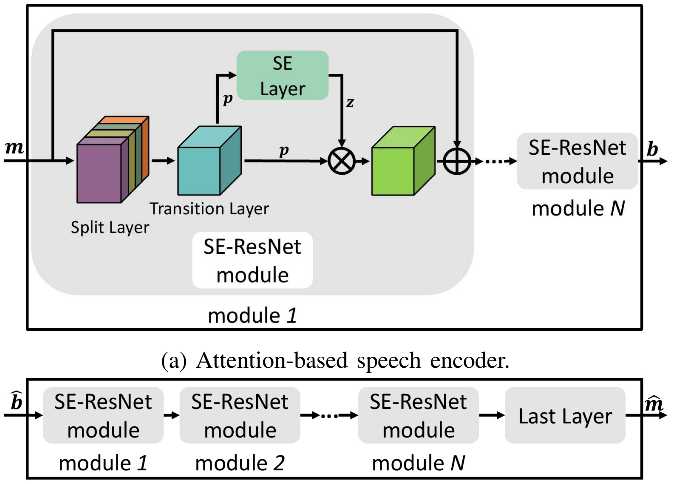


Fig. 2: The proposed system architecture for the speech semantic communication system.



(a) Attention-based speech encoder.

(b) Attention-based speech decoder.

Fig. 3: The proposed *speech encoder* and *speech decoder* based on SE-ResNet.

training task, the MSE loss converges until the loss is no longer decreasing. The number of SE-ResNet modules is an important hyperparameter, which aims to facilitate the good performance of the *speech encoder/decoder* and the reasonable training time. Moreover, the *channel encoder/decoder* is to mitigate the distortion and attenuation by the physical channels and in the channel layer, the noise, $\boldsymbol{W}$, is generated by a fixed SNR value.

After passing through the whole transceiver, the sample sequences set, $\boldsymbol{S}$, is recovered into $\widehat{\boldsymbol{S}}$, the size of $\boldsymbol{S}$ and $\widehat{\boldsymbol{S}}$ should be equal. Furthermore, the loss is computed at the end of the receiver according to (4) and the parameters are updated by (5). When the training stage is finished, the trained networks are obtained for testing.

2) Testing Stage: Based on the trained networks $\mathbf{T}_{\alpha}^{\mathcal{S}}(\cdot)$, $\mathbf{T}_{\beta}^{\mathcal{C}}(\cdot)$, $\mathbf{R}_{\chi}^{\mathcal{C}}(\cdot)$, and $\mathbf{R}_{\delta}^{\mathcal{S}}(\cdot)$ from the outputs of Algorithm 1, the testing algorithm of DeepSC-S is illustrated in Algorithm 2. Note that the speech sample sequences used for testing are different from that used for training.

As shown in Algorithm 2, the trained model under a fixed channel condition is employed to test the performance under various fading channels directly without model retraining. Note that transfer learning can be adopted as a promising technique to yield efficient retraining task when coping with dynamic environment [24]. However, in this work, we facilitate the model with strong adaptation.

Algorithm 1 Training algorithm of the proposed DeepSC-S

```
Initialization: initialize parameters $\boldsymbol{\theta}^{\mathcal{T}(\mathbf{0})}$ and $\boldsymbol{\theta}^{\mathcal{R}(\mathbf{0})}$, $i = 0$.
Input: Speech sample sequences $\boldsymbol{S}$ from speech dataset $\mathfrak{S}$, a fading channel $\boldsymbol{h}$, noise $\boldsymbol{w}$ generated under a fixed SNR value.
Framing $\boldsymbol{S}$ into $\boldsymbol{m}$ with trainable size.
while Stop criterion is not meet do
    $\mathbf{T}_{\alpha}^{\mathcal{S}}(\boldsymbol{m}) \to \boldsymbol{b}$.
    $\mathbf{T}_{\beta}^{\mathcal{C}}(\boldsymbol{b}) \to \boldsymbol{X}$.
    Transmit $\boldsymbol{X}$ over physical channel and receive $\boldsymbol{Y}$ via
    (2).
    $\mathbf{T}_{\chi}^{\mathcal{C}}(\boldsymbol{Y}) \to \widehat{\boldsymbol{b}}$.
    $\mathbf{T}_{\delta}^{\mathcal{S}}(\widehat{\boldsymbol{b}}) \to \widehat{\boldsymbol{m}}$.
    Deframing $\widehat{\boldsymbol{m}}$ into $\widehat{\boldsymbol{S}}$.
    Compute loss $\mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}}, \boldsymbol{\theta}^{\mathcal{R}})$ via (4).
    Update trainable parameters simultaneously via SGD:
        $\boldsymbol{\theta}^{\mathcal{T}(i+1)} \leftarrow \boldsymbol{\theta}^{\mathcal{T}(i)} - \eta\nabla_{\boldsymbol{\theta}^{\mathcal{T}(i)}}\mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}}, \boldsymbol{\theta}^{\mathcal{R}})$   (8)
        $\boldsymbol{\theta}^{\mathcal{R}(i+1)} \leftarrow \boldsymbol{\theta}^{\mathcal{R}(i)} - \eta\nabla_{\boldsymbol{\theta}^{\mathcal{R}(i)}}\mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}}, \boldsymbol{\theta}^{\mathcal{R}})$   (9)
    $i \leftarrow i + 1$.
end while
Output: Trained networks $\mathbf{T}_{\alpha}^{\mathcal{S}}(\cdot)$, $\mathbf{T}_{\beta}^{\mathcal{C}}(\cdot)$, $\mathbf{R}_{\chi}^{\mathcal{C}}(\cdot)$, and $\mathbf{R}_{\delta}^{\mathcal{S}}(\cdot)$.
```

V. Experiment and Numerical Results

In this section, we compare to the performance of the proposed DeepSC-S, the traditional communication systems and the system with an extra feature encoder for speech transmission under the AWGN channels, the Rayleigh channels, and the Rician channels, where the accurate CSI is assumed. The details of the adopted benchmarks will be introduced in part A of this section. Moreover, in order to facilitate DeepSC-S with good adaptation to practical environment, the DeepSC-S is tested over telephone systems and multimedia transmission systems, respectively. Note that the channel environment is modeled as a fading channel with a fixed SNR of 8 dB during the training stage and the testing channel set, $\mathcal{H}$, includes the AWGN channels, the Rayleigh channels, and the Rician channels.

In the whole experiment, we adopt the speech dataset from Edinburgh DataShare, which comprises more than 10,000 *.wav* files trainset and 800 *.wav* files testset with sampling rate 16KHz. In terms of the traditional telephone systems and multimedia transmission systems, the sampling rates for

Algorithm 2 Testing algorithm of the proposed DeepSC-S

```
Input: Speech sample sequences $\boldsymbol{S}$ from speech dataset $\mathfrak{S}$, trained networks $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\cdot)$, $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\cdot)$, $\mathbf{R}_{\boldsymbol{\chi}}^{\mathcal{C}}(\cdot)$, and $\mathbf{R}_{\boldsymbol{\delta}}^{\mathcal{S}}(\cdot)$, testing channel set $\mathcal{H}$, a wide range of SNR regime.
Framing $\boldsymbol{S}$ into $\boldsymbol{m}$ with trainable size.
for each channel condition $\boldsymbol{h}$ drawn from $\mathcal{H}$ do
    for each SNR value do
        generated Gaussian noise $\boldsymbol{w}$ under the SNR value.
        $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\boldsymbol{m}) \to \boldsymbol{b}$.
        $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\boldsymbol{b}) \to \boldsymbol{X}$.
        Transmit $\boldsymbol{X}$ over physical channel and receive $\boldsymbol{Y}$
        via (2).
        $\mathbf{T}_{\boldsymbol{\chi}}^{\mathcal{C}}(\boldsymbol{Y}) \to \widehat{\boldsymbol{b}}$.
        $\mathbf{T}_{\boldsymbol{\delta}}^{\mathcal{S}}(\widehat{\boldsymbol{b}}) \to \widehat{\boldsymbol{m}}$.
        Deframing $\widehat{\boldsymbol{m}}$ into $\widehat{\boldsymbol{S}}$.
    end for
end for
Output: Recovered speech sample sequences, $\widehat{\boldsymbol{S}}$, under different fading channels and various SNR values.
```

speech signals are 8KHz and 44.1KHz, respectively. Thus, for the experiment regarding telephone systems, the input samples are down-sampled to 8KHz and regarding multimedia communications, the input samples are up-sampled to 44.1KHz. Note that the number of speech samples in different *.wav* is inconsistent. In the simulation, we fix $W = 16,384$, and each sample sequence in $\boldsymbol{m}$ consists of frames $F = 128$ with the frame length $L = 128$.

A. Benchmark Model

We use the following three different beachmarks. For the traditional communications, speech transmission over telephone systems has lower accuracy requirements compare to multimedia transmission systems. For instance, audio signals in video required to be extremely clear, but the background noise and echo occur when speaking over the phone.

1) Benchmark 1: According to ITU-T G.711 standard, 64 Kbps pulse code modulation (PCM) is recommended for speech source coding in telephone systems with $2^8 = 256$ quantization levels [35]. Moreover, 16-bits PCM is adopted in our work for speech transmission in multimedia transmission systems with $2^{16} = 65,536$ quantization levels. Note that A-law PCM and uniform PCM are adopted in telephone systems and multimedia transmission systems, respectively. For the channel coding, turbo codes with soft output Viterbi algorithm (SOVA) is considered to improve the performance of error detection and correction at the receiver [36], in which the coding rate is 1/3, the block length is 512, and the number of decoding iterations is 5. In addition, to make the number of transmitted symbols in the traditional systems is same as that in DeepSC-S, 64-QAM is adopted in the benchmark for the modulation. The details for the the typical communication systems for the two different transmission applications are summarized in Table I.

TABLE I: Parameters settings in the traditional communication systems.

	Telephone Systems	**Multimedia Systems**
Sample rate	8KHz	44.1KHz
Signal Length	16384	16384
Number of frames	128	128
Frame length	128	128
Source coding	8-bits PCM	16-bits PCM
Channel coding	Turbo codes	Turbo codes
Modulation	64-QAM	64-QAM

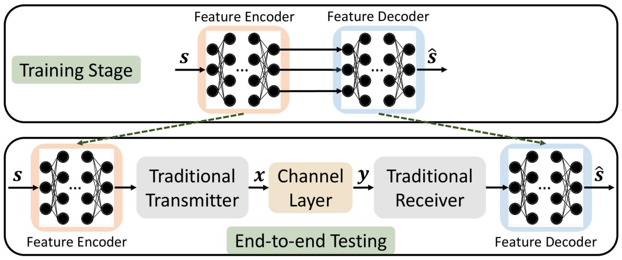


Fig. 4: The benchmark model by combing a feature encoder with the transmission systems.

2) Benchmark 2: The second benchmark combines a feature encoder with the traditional model, named a semi-traditional communication system, as shown in Fig. 4. From the figure, at the training stage, the feature encoder takes the speech samples, $\boldsymbol{s}$, as the input and the output is fed into the feature decoder directly. The received signal is converted into the speech information, $\widehat{\boldsymbol{s}}$, by the feature decoder. Based on signals $\boldsymbol{s}$ and $\widehat{\boldsymbol{s}}$, the MSE loss is computed at the end of the receiver, thus, the trainable parameters of the feature encoder and the feature decoder are updated via SGD at the same time.

For the end-to-end testing, the pre-trained feature leaning system is split into the feature encoder and the feature decoder, which are placed before the traditional transmitter and after the traditional receiver, respectively. Note that the signal processing blocks of the traditional communication system are same as the settings as shown in Table I. During the training stage, the feature encoder and the feature decoder are treated as the extraction and recovery operations without considering communication problems. During the end-to-end testing stage, the semi-traditional system is aimed to yield efficient transmission as well as to mitigate the channel effects. The network settings of the semi-traditional system are shown as Table II.

TABLE II: Parameters settings of the semi-traditional system.

	Layer Name	**Kernels**	**Activation**
Feature Encoder	6×CNN layer	6×32	Relu
	CNN layer	1	None
Feature Decoder	6×CNN layer	6×32	Relu
	CNN layer	1	None
Learning Rate	η	0.001	None

3) Benchmark 3: In order to emphasize the improvement of the attention mechanism, we combine a semantic communication system without attention into simulation based on 2D CNN module, named a CNN-based system. The network settings of the CNN-based system are shown as Table III.

TABLE III: Parameters settings of the CNN-based system for telephone systems.

	Layer Name	Kernels	Activation
Transmitter	6×CNN modules	6×32	Relu
	CNN layer	8	Relu
Receiver	CNN layer	8	Relu
	6×CNN modules	6×32	Relu
	Last layer (CNN)	1	None
Learning Rate	η	0.001	None

B. Experiments over Telephone Systems

In this experiment, we first investigate a robust system to work on various channel conditions while training DeepSC-S under the fixed channel condition, and then testing the MSE loss via the trained model under all adopted fading channels. Besides, we test the SDR and PESQ under DeepSC-S, the traditional system, the semi-traditional systems, and the CNN-based system for speech transmission over telephones systems. Particularly, the number of the SE-ResNet modules in the *speech encoder/decoder* is 6 and the number of the 2D CNN modules in the *channel encoder/decoder* is 1, which includes 8 kernels. For each SE-ResNet module, the number of kernels in the *split* layer and the *transition* layer is 32. The learning rate is set as 0.001. The network setting of the proposed DeepSC-S are shown as Table IV.

TABLE IV: Parameters settings of the proposed DeepSC-S for telephone systems.

	Layer Name	Kernels	Activation
Transmitter	6×SE-ResNet	6×32	Relu
	CNN layer	8	Relu
Receiver	CNN layer	8	Relu
	6×SE-ResNet	6×32	Relu
	Last layer (CNN)	1	None
Learning Rate	η	0.001	None

As shown in Fig. 5a, in terms of the MSE loss tested under the AWGN channels, DeepSC-S trained under the AWGN channels outperforms the model trained under the Rayleigh channels and the Rician channels when SNRs are higher than around 6 dB. However, it has higher MSE loss values in the low SNR regime. Besides, according to Fig. 5b, DeepSC-S trained under the AWGN channels performs quite poor in terms of MSE loss when testing for Rayleigh channels. Furthermore, Fig. 5c shows the model trained under the three adopted channels can achieve MSE loss values under 9×10^{-7} when testing under the Rician channels. Therefore, DeepSC-S trained under the Rician channels is considered as a robust model that is capable of coping with various channel environments.

Fig. 6 tests the SDR performance between the traditional communication systems, the semi-traditional system, the CNN-based system, and the proposed DeepSC-S under the AWGN channels, the Rayleigh channels, and the Rician channels. From the figure, the semi-traditional system yields higher SDR score than the traditional one under all tested channel environments while it performs unreliable when SNRs are low. Besides, the CNN-based system and DeepSC-S achieve better SDR than the the semi-traditional system and the traditional system under the Rayleigh channels and the Rician channels, as well as the AWGN channels over the most ranges of tested SNRs. In addition, DeepSC-S performs steadily when coping with different channels and SNRs, however, for the semi-traditional system and the traditional system, the performances are quite poor under dynamic channel conditions, especially in the low SNR regime. Moreover, due to the attention mechanism, SE-ResNet, the proposed DeepSC-S achieves higher SDR score than the CNN-based system under all adopted SNRs and fading channels, which proves the effectiveness of the DeepSC-S.

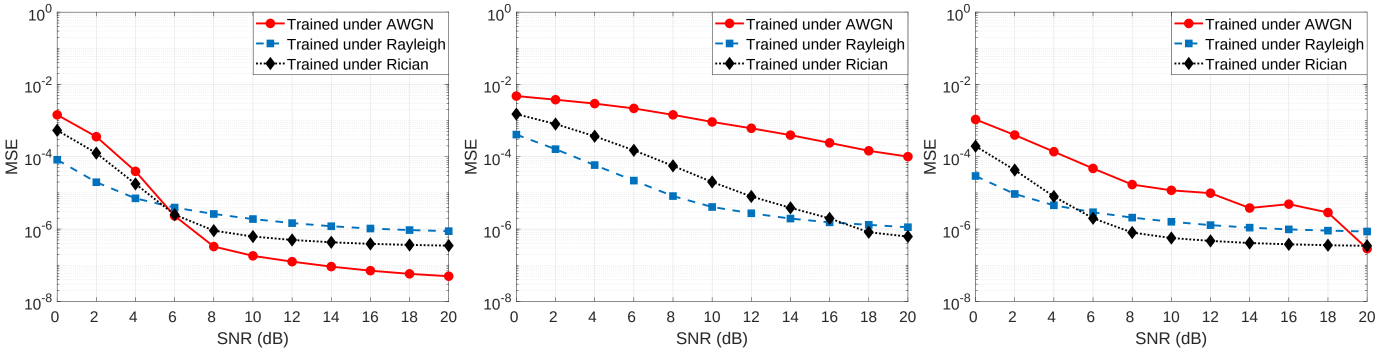
The PESQ score comparison is shown in Fig. 7. From the figure, the CNN-based system and DeepSC-S provide high quality speech recovery and outperform the semi-traditional system and the traditional system under various fading channels and SNRs. Moreover, similar to the results of SDR, DeepSC-S obtains good PESQ when coping with channel variation while the traditional one provides poor scores in the low SNR regime. DeepSC-S also achieves higher score than the CNN-based system under all adopted channel conditions. Based on the simulated results, the proposed DeepSC-S is able to yield better speech transmission for the telephone systems under complicated communication scenarios than the traditional systems, especially in the low SNR regime.

C. Experiments over Multimedia Transmission Systems

In this part, we present the SDR and PESQ performance comparison between DeepSC-S and the traditional systems for speech signals transmission for multimedia applications, as well as the CNN-based system similar to the telephone communications experiment. The NN network settings of the CNN-based system and the proposed DeepSC-S are similar to Table III and Table IV, respectively, but the kernels of CNN layer at the transmitter and the receiver in both CNN-based system and DeepSC-S are 16.

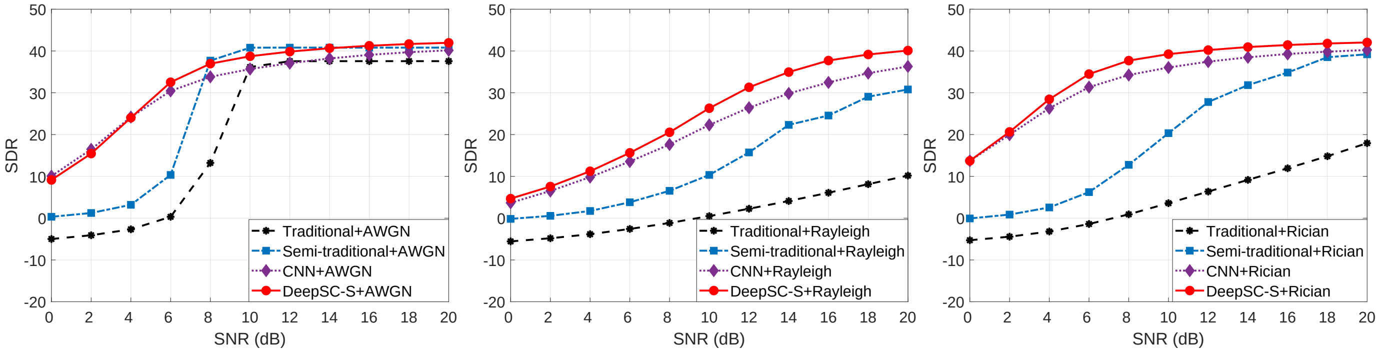
Fig. 8 depicts the SDR performance comparison for multimedia communications among the traditional system, the semi-traditional system, the CNN-based system, and the proposed DeepSC-S under the AWGN channels and the Rician channels. For the traditional system under the AWGN channels, the SDR score shows sharp increasing when the SNR is over 8 dB, and it achieves SDR scores over 80 in high SNRs due to the high PCM quantization accuracy. However, DeepSC-S can reach universal strong SDR values for all tested SNRs and fading channels. Moreover, DeepSC-S outperforms the semi-traditional system and the traditional system under the Rician channels, as well as the AWGN channels in the low SNR regime. Furthermore, DeepSC-S has higher SDR score than the CNN-based system because the SE-ResNet module is utilized to learn and extract the essential information.

The simulation result of PESQ for multimedia communications is illustrated in Fig. 9. From the figure, the proposed DeepSC-S outperforms the semi-traditional system and the the traditional system under the Rician channels with any tested SNRs as well as the AWGN channels with low SNR values. Moreover, similar to the results of SDR, the proposed DeepSC-S achieves higher PESQ score than the CNN-based system



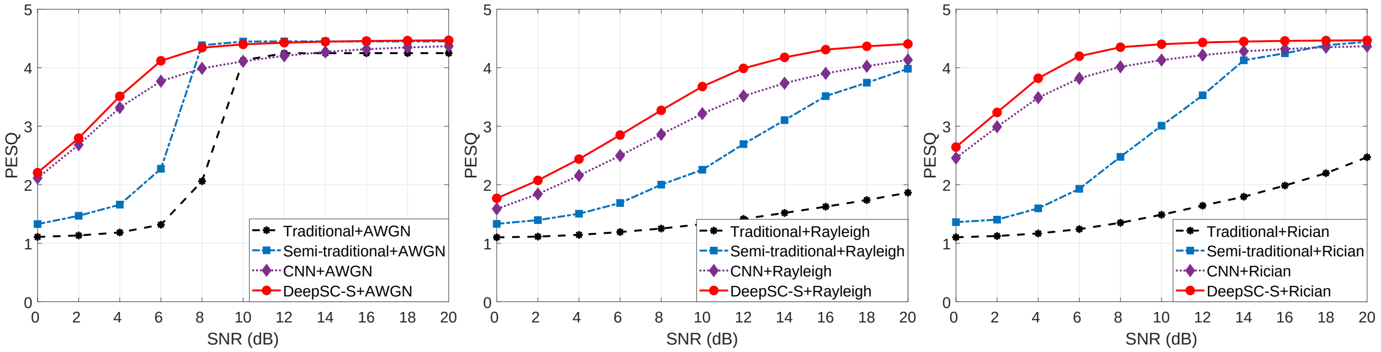
(a) AWGN channels (b) Rayleigh channels (c) Rician channels

Fig. 5: MSE loss tested for (a) AWGN, (b) Rayleigh, and (c) Rician channels with the models trained under various channels.



(a) AWGN channels (b) Rayleigh channels (c) Rician channels

Fig. 6: SDR score versus SNR for speech based telephone communications with the traditional system, the semi-traditional system, the CNN system, and the proposed DeepSC-S for (a) AWGN channels, (b) Rayleigh channels, (c) Rician channels.



(a) AWGN channels (b) Rayleigh channels (c) Rician channels

Fig. 7: PESQ score versus SNR for speech based telephone communications with the traditional system, the semi-traditional system, the CNN system, and the proposed DeepSC-S for (a) AWGN channels, (b) Rayleigh channels, (c) Rician channels.

under all adopted channel conditions. Thus, it is believed that the investigated DeepSC-S is with greater adaptability than the traditional system for speech based multimedia communications when coping with channel variation.

VI. Conclusion

In this article, we investigate a DL-enabled semantic communication system for speech transmission, named DeepSC-S, which achieves more efficient transmission than the traditional approaches by utilizing the semantic information of speech signals. Particularly, we jointly design the *speech encoder/decoder* and the *channel encoder/decoder* to learn and extract the speech features, as well as to mitigate the channel distortion and attenuation for practical communication scenarios. Additionally, an attention mechanism based on squeeze-and-excitation (SE) networks is utilized to improve the recovery accuracy by minimizing the mean-square error of the speech signals. Moreover, in order to enable DeepSC-S working well over various physical channels, a DeepSC-S model with strong robustness to channel variations is in-

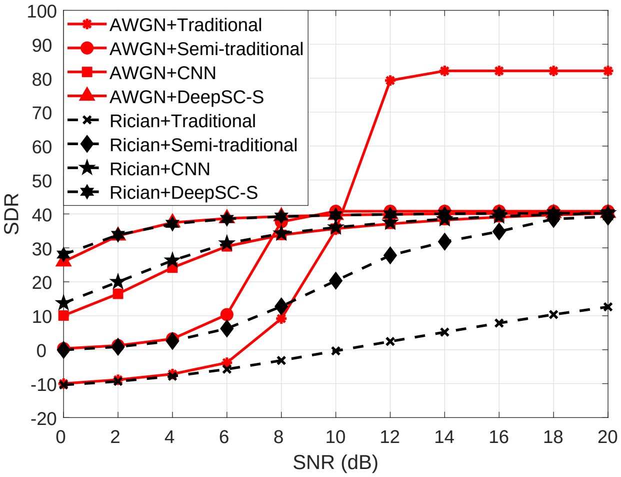


Fig. 8: SDR score versus SNR for speech based multimedia communications with the traditional system, the semi-traditional system, the CNN system, and the proposed DeepSC-S for AWGN channels and Rician channels.

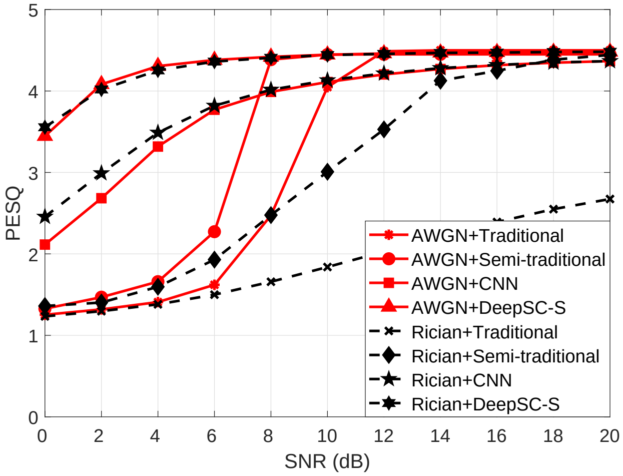


Fig. 9: PESQ score versus SNR for speech based multimedia communications with the traditional system, the semi-traditional system, the CNN system, and the proposed DeepSC-S for (a) AWGN channels, (b) Rayleigh channels, (c) Rician channels.

vestigated. The proposed DeepSC-S is investigated under the telephone systems and the multimedia transmission systems for verifying the system adaptation. Simulation results demonstrated that DeepSC-S outperforms the traditional communication systems as well as the semi-traditional system with an extra feature encoder, especially when the SNR is low. Hence, our proposed DeepSC-S is a promising candidate for speech semantic communication systems.

References

[1] Z. Qin, H. Ye, G. Y. Li, and B.-H. F. Juang, “Deep learning in physical layer communications,” *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[2] Z. Qin, G. Y. Li, and H. Ye, “Federated Learning and Wireless Communications,” https://arxiv.org/abs/2005.05265, May. 2020.

[3] T. Gruber, S. Cammerer, J. Hoydis, and S. T. Brink, “On deep learning-based channel decoding,” in *Proc. IEEE 51st Annu. Conf. Inf. Sci. Syst. (CISS),* Baltimore, MD, USA, Mar. 2017, pp. 1–6.

[4] H. Ye, G. Y. Li, and B.-H. F. Juang, “Power of deep learning for channel estimation and signal detection in OFDM systems,” *IEEE Wireless Commun. Lett.,* vol. 7, no. 1, pp. 114-117, Feb. 2018.

[5] N. Samuel, T. Diskin, and A. Wiesel, “Deep MIMO detection,” in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.,* Sapporo, Japan, Dec. 2017, pp. 690–694.

[6] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, “Learning to optimize: Training deep neural networks for interference management,” *IEEE Trans. Signal Process.,* vol. 66, no. 20, pp. 5438-5453, Oct. 2018.

[7] L. Liang, H. Ye, G. Yu, and G. Y. Li, “Deep-learning-based wireless resource allocation with application to vehicular networks,” *Proc. IEEE,* vol. 108, no. 2, pp. 341–356, Feb. 2020.

[8] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communications.* The University of Illinois Press, 1949.

[9] R. Carnap and Y. Bar-Hillel, *An Outline of a Theory of Semantic Information.* RLE Technical Reports 247, Research Laboratory of Electronics, Massachusetts Institute of Technology., Cambridge MA, Oct. 1952.

[10] J. Bao, P. Basu, M. Dean, C. Partridge, A. Swami, W. Leland, and J. A. Hendler, “Towards a theory of semantic communication,” in *Proc. IEEE Netw. Sci. Workshop,* West Point, NY, USA, Jun. 2011, pp. 110–117.

[11] P. Basu, J. Bao, M. Dean, and J. Hendler, “Preserving quality of information by using semantic relationships,” *Pervasive Mob. Comput.,* vol. 11, pp. 188–202, Apr. 2014.

[12] T. O’Shea and J. Hoydis, “An introduction to deep learning for the physical layer,” *IEEE Trans. Cogn. Commun. Netw.,* vol. 3, no. 4, pp. 563–575, Dec. 2017.

[13] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” *IEEE Trans. Audio, Speech, Language Process.,* vol. 20, no. 1, pp. 33–42, Jan. 2012.

[14] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR),* Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[15] M. Kim, W. Lee, and D-H. Cho, “A novel PAPR reduction scheme for OFDM system based on deep learning,” *IEEE Commun. Lett.,* vol. 22, no. 3, pp. 510–513, Mar. 2018.

[16] A. Felix, S. Cammerer, S. Dörner, J. Hoydis, and S. T. Brink, “OFDM autoencoder for end-to-end learning of communications systems,” in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.,* Kalamata, Greece, Jun. 2018, pp. 1-5.

[17] T. J. O’Shea, T. Erpek, and T. C. Clancy, “Physical layer deep learning of encodings for the MIMO fading channel,” in *Proc. Annu. Allerton Conf. Commun., Control, Comput. (Allerton),* Monticello, IL, USA, Oct. 2017, pp. 76–80.

[18] H. He, C. Wen, S. Jin, and G. Y. Li, “Model-driven deep learning for MIMO detection,” *IEEE Trans. Signal Process.,* vol. 68, pp. 1702–1715, Feb. 2020.

[19] F. A. Aoudia, and J. Hoydis, “Model-free training of end-to-end communication systems,” *IEEE J. Sel. Areas Commun.,* vol. 37, no. 11, pp. 2503-2516, Nov. 2019.

[20] H. Ye, L. Liang, G. Y. Li, and B.-H. Juang, “Deep learning based end-to-end wireless communication systems with conditional GAN as unknown channel,” *IEEE Trans. Wireless Commun.,* vol. 19, no. 5, pp. 3133-3143, May. 2020.

[21] S. J. Pan, and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.,* vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[22] S. Park, O. Simeone, and J. Kang, “Meta-learning to communicate: Fast end-to-end training for fading channels,” in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP),* Barcelona, Spain, May. 2020, pp. 5075-5079.

[23] B. Guler, A. Yener, and A. Swami, “The semantic communication game,” *IEEE Trans. Cogn. Commun. Netw.,* vol. 4, no. 4, pp. 787–802, Sep. 2018.

[24] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep learning enabled semantic communication systems,” https://arxiv.org/abs/2006.10685, May. 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances Neural Info. Process. Syst. (NIPS’17),* Long Beach, CA, USA. Dec. 2017, pp. 5998–6008.

[26] H. Xie, and Z. Qin, “A lite distributed semantic communication system for Internet of Things,” *IEEE J. Sel. Areas Commun.* vol. 39, no. 1, pp. 142–153, Jan. 2021.

[27] E. Bourtsoulatze, D. Burth Kurka, and D. Gündüz, “Deep joint source-channel coding for wireless image transmission,” *IEEE Trans. Cogn. Commun. Netw.,* vol. 5, no. 3, pp. 567–579, Sept. 2019.

[28] D. B. Kurka and D. Gündüz, “Deepjscc-f: Deep joint source-channel coding of images with feedback,” *IEEE J. Sel. Areas Inf. Theory,* vol. 1, no. 1, pp. 178–193, Apr. 2020.

[29] C. Lee, J. Lin, P. Chen, and Y. Chang, “Deep learning-constructed joint transmission-recognition for Internet of Things,” *IEEE Access,* vol. 7, pp. 76547–76561, Jun. 2019.

[30] M. Jankowski, D. Gündüz, and K. Mikolajczyk, “Joint device-edge inference over wireless links with pruning,” in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.,* Atlanta, GA, USA, Aug. 2020, pp. 1–5.

[31] M. Jankowski, D. Gündüz, and K. Mikolajczyk, “Wireless image retrieval at the edge,” *IEEE J. Sel. Areas Commun.* vol. 39, no. 1, pp. 89–100, Jan. 2021.

[32] E. Vincent, R. Gribonval, and C. Fevotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Language Process.,* vol. 14, no. 4, pp. 1462–1469, Jun. 2006.

[33] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.. Proc. (ICASSP),* Salt Lake City, UT, USA, May. 2001, pp. 749–752.

[34] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,* ITU-T recommendation P.862, Mar. 2018.

[35] R. Cox, “Three new speech coders from the ITU cover a range of applications,” *IEEE Commun. Mag.,* vol. 35, no. 9, pp. 40–47, Sept. 1997.

[36] Y. Wu and B. Woerner, “The influence of quantization and fixed point arithmetic upon the BER performance of turbo codes,” in *Proc. IEEE Veh. Technol. Conf. (VTC),* Houston, TX, USA, May. 1999, pp. 1683–1687.