This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

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 semantic information. In this paper, 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 or symbol level. Particularly, we design a deep learning (DL)-enabled semantic communication system for speech signals, named DeepSC-S. In order to improve the recovery accuracy of speech signals, especially for the essential information, DeepSC-S is developed based on an attention mechanism by utilizing a squeeze-and-excitation (SE) network. The motivation behind the attention mechanism is to identify the essential speech information by providing higher weights to them when training the neural network. Moreover, in order to facilitate the proposed DeepSC-S for dynamic channel environments, we find a general model to cope with various channel conditions without retraining. Furthermore, we investigate DeepSC-S in telephone systems as well as multimedia transmission systems to verify the model adaptation in practice. The simulation results demonstrate that our proposed DeepSC-S outperforms the traditional communications in both cases in terms of the speech signals metrics, such as signal-to-distortion ration and perceptual evaluation of speech distortion. Besides, DeepSC-S is more robust to channel variations, especially in the low signal-to-noise (SNR) regime.**

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

I. Introduction

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], [3]. Particularly, DL has shown its great potentials to solve the existing technical problems in both physical layer communications [4]–[6] and wireless resource allocations [7], [8].

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 a neural network (NN) to represent the intelligent transceiver. However, even if the DL-enabled communication systems yield better performance and/or lower complexity for some scenarios and conditions, their state-of-the-art models mainly focus on performance improvement at the bit or symbol level, which usually takes bit-error rate (BER) or symbol-error rate (SER) as the performance metric. 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) with the system capacity approaching Shannon limit.

Shannon and Weaver [9] 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 wide deployment of intelligent IoT applications, semantic-irrelative communications are no longer ideal as they transmit bit sequences, which contains information that could be relevant or not to the intelligent tasks at the receiver. Moreover, in the typical communication systems, the generated data is more than required, which limits the number of devices to be covered by the same network. Motivated by this, researchers are dedicating to develop a new system to process and exchange semantic information for more efficient communications.

Semantic theory takes into account the meaning and veracity of source information because they can be both informative and factual [10], which facilitates the semantic communication systems to transmit only the semantic information at the transmitter and to recover information at the receiver via minimizing the semantic error instead of BER/SER. Nevertheless, the exploration of semantic communications has gone through decades of stagnation since it was first identified because of the limitation of some fundamental problems, e.g., lack of mathematical model for semantic information. The semantic information refers to the information relevant to the transmission goal at the receiver, however, even the most cutting-edge work cannot define the semantic information or semantic features by a precise mathematical formula. Moreover, the semantic information varies for different transmission purposes, which could be in various formats, e.g., age of information [11], or more complicated semantic features.

Semantic data can be compressed to a proper size for transmission by using a lossless method [12], which utilizes the semantic relationship between different messages, while the traditional lossless source coding is to represent a signal

Part of the work has been presented in [1], which has been accepted by IEEE ICC 2021.

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)(Corresponding author: Zhijin Qin).

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

with the minimum number of binary bits by exploring the dependencies or statistical properties of input signals. In addition, an end-to-end (E2E) communication system has been developed [13] in order to address the bottlenecks in traditional block-wise communication systems. Inspired by this, different types of sources have been considered in recent investigations on E2E semantic communication systems, which mainly focuses on the image and text transmission [14]–[21]. The investigation on semantic communication for speech signals transmission is still missed.

A semantic communication system with different inputs is shown in Fig. 1, which only transmits the semantic features highly relevant to the transmission task at the receiver. While the transmission tasks could be either the source message recovery or more intelligent tasks. For example, in speech signal processing, one intelligent task is to convert speech signals into text information, e.g., automatic speech recognition (ASR), but do not care about the characteristics of speech signals, e.g., the speaking speed and tone [22]. The core of ASR is to map each phoneme into a single alphabet, and then concatenate all alphabets into an understandable word sequence via a language model. In this case, the extracted semantic features only contain the text characteristics while the other features will not be transmitted by the transmitter. As a result, the network traffic is reduced significantly without performance degradation.

Without loss of generality, we consider a general model of semantic communication to restore the speech sources in this work, which could be extended to serve a specific intelligent task easily in the future. The intuition behind this work is to recover speech signals based on DL technique, which includes the recovery of speech characteristics. However, 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 extra operations capture the unique features of speech signals but runs counter to the motivation of intelligence. Therefore, a DL-enabled semantic communication system by learning semantic information directly from the raw speech signals is of great interest and importance.

In this paper, we propose a DL-enabled semantic communication system for speech signals, named DeepSC-S, 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:

- The DeepSC-S is first developed, which treats the transmitter and the receiver as two NNs. The joint semantic-channel coding is developed to deal with source distortions and channel effects.
- Particularly, a squeeze-and-excitation (SE) network [23] is employed in the developed DeepSC-S 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 an SE network, DeepSC-S improves the accuracy of signal recovery.
- Moreover, a trained model with high robustness to channel variations is developed by training DeepSC-S under a fixed channel condition, and then facilitating it with good performance when coping with different testing channel environments.
- 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 compared with the traditional approaches to prove its superiority, especially in the low SNR regime.

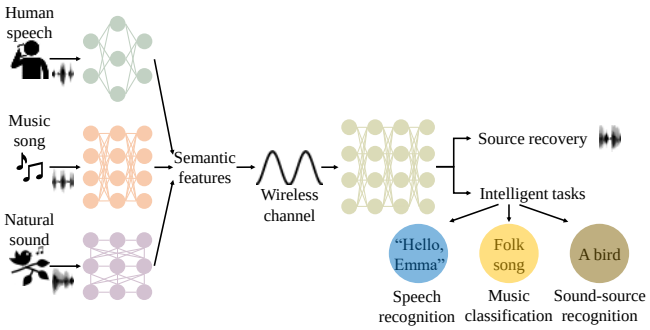


Fig. 1: A semantic communication system with different inputs.

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 are presented. Simulation results are discussed in Section V. Section VI draws conclusions.

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

II. Related Work

This section first introduce the related work on E2E communication systems to address the challenges in traditional systems. Then, we discuss the state-of-the-art semantic communications.

A. End-to-End Communication Systems

The DL-enabled E2E communication systems have achieved extremely competitive block-error rate (BLER) performance in comparison to the traditional block-wise communication systems in various scenarios [13]. In addition, it has shown great potentials in processing complicated communication tasks. For examples, the E2E learning has been employed

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

in orthogonal frequency division multiplexing (OFDM) systems [24], [25], as well as in multiple-input multiple-output (MIMO) systems [26], [27]. Besides, channel estimation is a challenging problem in the DL-enabled E2E systems. In [28], reinforcement learning (RL) has been adopted to estimate the channel state information (CSI) through treating the channel layer and the receiver as the *environment*. However, it requires a reliable channel to feedback the losses from the receiver to the transmitter during the training phase. Another novel channel agnostic solution has been proposed in [29], which replaces the realistic channel with a deep neural network (DNN) by exploiting a conditional generative adversarial network (GAN).

Due to the complexity of NN training, a system with high training efficiency and low energy consumption is more than desired to make the E2E system applicable in practice. Transfer learning is 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 [30]. Another appealing solution is to obtain a trained model yielding expected performance via the stochastic gradient descent (SGD) with a small number of iterations. Particularly, based on model-agnostic meta-learning (MAML), an E2E communication system has been developed in [31], which finds the initial NN parameters to achieve fast convergence for various channel conditions.

B. Semantic Communications

An initial research on semantic communication systems for text information was developed [14], which mitigates the semantic error by integrating the semantic inference and the physical layer communications to optimize the whole transceiver. However, such a text-based semantic communication system only measures the semantic error at the word level instead of the sentence level. Thus, a further investigation on semantic text transmission, named DeepSC, has been developed [15] to deal with the semantic error at the sentence level with various length. Powered by the Transformer [32], the semantic encoder and the channel encoder are co-designed to minimize the semantic error and to improve the system capacity. Moreover, the increasing deployment of smart IoT applications requires IoT devices to be capable of dealing with more complicated tasks, 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 [16], has been further proposed to address the challenge of IoT devices by pruning and quantizing NN parameters. By doing so, the size of the trained model as well as the communication cost between the IoT devices and the server are reduced significantly, which makes it more suitable for IoT applications.

In semantic communications for image information, a DL-enabled semantic communication system has been developed [17], which employs a convolutional neural network (CNN) at the transmitter to jointly design the source-channel coding. Besides, an image transmission system has been investigated to improve the accuracy of image reconstruction [18], which backpropagates the channel output in order to generate a weight vector to the NN at the transmitter. In addition to perform the typical image reconstruction, more intelligent tasks have been considered. Particularly, an application based on a joint source-channel coding (JSCC) model to retrieve image has been proposed [19], which aims to reduce the transmission latency for IoT devices by achieving retrieval-oriented image compression. Besides, a joint image transmission-recognition system has been developed [20] for IoT applications, which has the superior recognition accuracy than the traditional approaches but requires the affordable computation resource. Moreover, by exploiting NN compression techniques, a deep JSCC [21] has been developed to perform image classification at the edge sever, which facilitates the IoT devices to process images with low computational complexity and to reduce the required transmission bandwidth. Note that the concept of semantic communications was not clearly stated by the authors in the aforementioned work on image transmission. However, we treat them as the pioneering works in the area as they share the spirit of semantic communications by extracting and transmitting the relevant information from the source for serving the transmission goals at the receiver.

III. System Model

In this section, we first introduce the considered system model, then the performance metrics are presented. The considered system transmits the original speech signals via a NN-based speech semantic communication system, which comprises two major tasks: i) semantic information learning and extracting from speech signals; ii) mitigating the effects of wireless channels. 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 speech signals and to outperform the traditional approaches when coping with complicated channel conditions.

A. Transmitter

The proposed system model is shown in Fig. 2. 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 wth 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 shown in Fig. 2, the transmitter consists of two individual components: the *semantic encoder* and the *channel encoder*, each component is implemented by an independent NN. Denote the NN parameters of the *semantic encoder* and the *channel encoder* as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Then the encoded symbol sequence, $\boldsymbol{x}$, can be expressed as

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

where $\mathbf{T}^{\mathcal{S}}_{\boldsymbol{\alpha}}(\cdot)$ and $\mathbf{T}^{\mathcal{C}}_{\boldsymbol{\beta}}(\cdot)$ indicate the *semantic 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 $\boldsymbol{\theta}^{\mathcal{T}} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$.

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

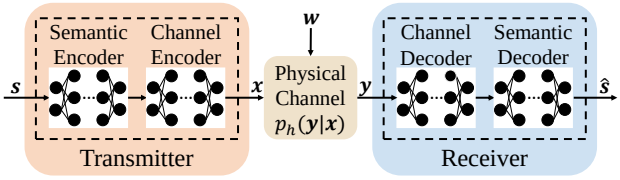


Fig. 2: The model structure of DL-enabled speech semantic communication system.

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 transmit power constraint $\mathbb{E}\left\|\boldsymbol{x}\right\|^2 = 1$.

The whole transceiver in Fig. 2 is designed for a single communication link. 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.

B. Receiver

Similar to the transmitter, the receiver also consists of two cascaded parts, including the *channel decoder* and the *semantic decoder*. The *channel decoder* is to mitigate the channel distortion and attenuation, and the *semantic decoder* recovers speech signals based on the learned and extracted speech semantic features. Denote the NN parameters of the *channel decoder* and the *semantic decoder* as $\boldsymbol{\chi}$ and $\boldsymbol{\delta}$, respectively. As depicted in Fig. 2, 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 *semantic decoder* w.r.t. parameters $\boldsymbol{\chi}$ and $\boldsymbol{\delta}$, respectively. Denote the NN parameters 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 signals, which causes two challenges. The first one is the design of intelligent *semantic encoder/decoder*, which utilizes 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. 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 *semantic encoder* and the *channel encoder*, and recovered 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

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

where s_w and $\widehat{s}_w$ indicate the wth element of 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 (4). It is worth to mention that the *semantic 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 parameters 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, which iteratively updates the parameters $\boldsymbol{\theta}$ as 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.

C. Performance Metrics

In our model, the system is committed to reconstruct the raw speech signals. Hence, the signal-to-distortion ration (SDR) [33] 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

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

The higher SDR represents that 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.

Furthermore, the good recovery of speech signals is intuitively manifested by a satisfactory listening experience of the recovered speech signals, e.g., no latency and background noise. The perceptual evaluation of speech distortion (PESQ) [34] is adopted in the International Telecommunication Union (ITU-T) recommendation P.862 [35], which is a good candidate for evaluating the quality of speech signals under various conditions, e.g., background noise, analog filtering, and variable delay, by scoring the quality from -0.5 to 4.5. The PESQ score is obtained by multiple operations, e.g., level align, time align and equalise, and disturbance processing [34]. In this work, we adopt an integrated open source PESQ assessment model developed by the ITU-T [35], which is able to evaluate the PESQ score in few milliseconds.

IV. Proposed Semantic Communication System for Speech Signals

To address the aforementioned challenges, we design a DL-enabled semantic communication system for speech transmission, named DeepSC-S. Specifically, an attention-based two-dimension (2D) CNN is used for the *semantic encoder/decoder*

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

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. 3, the input of the proposed DeepSC-S, denoted as $\boldsymbol{S} \in \Re^{B\times W}$, is a set of speech sample sequences, $\boldsymbol{s}$, which is drawn from the speech dataset, $\mathfrak{S}$, and B is the batch size. $\mathfrak{S}$ consists of considerable speech sequences, which are collected by recording the speakings from different persons. The input sample sequences, $\boldsymbol{S}$, are framed into $\boldsymbol{m} \in \Re^{B\times F\times L}$ for training before passing through an attention-based encoder, i.e., the *semantic 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 *semantic 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 *semantic encoder* are presented 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 signal, $\boldsymbol{Y}$, 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 *semantic 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. 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. Semantic Encoder and Decoder

It is intuitive that speech signals at the silent time-slots carry no information, while speech signals at the speaking time-slots carry the essential information. Therefore, signals with zero magnitudes should be ignored, while signals with higher magnitudes should be processed with more attention. Moreover, at the speaking time-slots, people will speak loudly to emphasize the essential information. Correspondingly, the magnitudes of speech signals will increase abruptly. Similarly, people usually speak slowly to state the incomprehensible information, which results in a sudden drop on the frequency of corresponding speech signals. The essential information refers to the features that exist at the speaking time-slots, however, these features are hard or impossible to be captured by mathematical formula, which makes it difficult to measure the feature difference between the input speech signals and the recovered ones straightforwardly, or integrate them into loss function.

Inspired by this, we propose the DeepSC-S including the *semantic encoder* and the *semantic decoder* based on an attention mechanism, named SE-ResNet. In this work, SE-ResNet is utilized to learn and extract the essential information of speech signals. Particularly, SE-ResNet is capable to identify the essential information by assigning higher weights for them during the NN training process.

As shown in Fig. 4, for the SE-ResNet, a *split* layer takes $\boldsymbol{m}$ as the input and produces multiples blocks, besides, all these blocks are concatenated. Then a *transition* layer is utilized to reduce the dimension of these 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 in $\boldsymbol{p}$. The output of the SE layer, $\boldsymbol{z} \in \Re^{1\times 1\times C}$, includes C number of scale coefficients, which is utilized to scale the importance of the extracted features in $\boldsymbol{p}$, i.e., the features corresponding to the essential information in $\boldsymbol{p}$ are multiplied by the high scale coefficients in $\boldsymbol{z}$. By doing so, the weights of $\boldsymbol{m}$ are reassigned, i.e., the weights corresponding to the essential speech information are paid more attention. Note that the SE layer is 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 the essential information will improve, however, it also increases the computational cost. Therefore, a trade-off between the learning performance and complexity should be considered during the training phase. 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 SE-ResNet module, as shown in Fig. 4.

Particularly, the *semantic encoder* is comprised by multiple SE-ResNet modules to convert input $\boldsymbol{m}$ into $\boldsymbol{b}$, corresponding to Fig. 3. For the *semantic decoder*, in addition to several SE-ResNet modules, the *last layer*, including a 2D CNN module with a single *filter*, is utilized to reduce the size of output $\widehat{\boldsymbol{m}}$, because the size 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 the 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 recover the speech signals and make it to work well under various fading channels and a wide SNR regime.

1) Training Stage: According to Fig. 3, 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

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

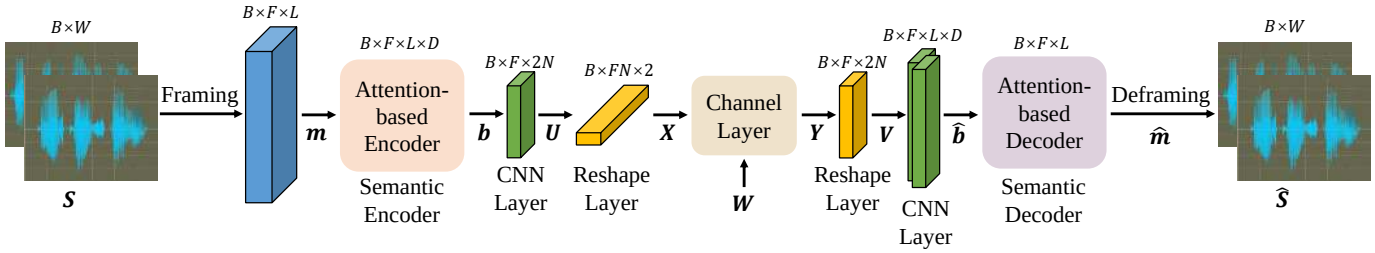
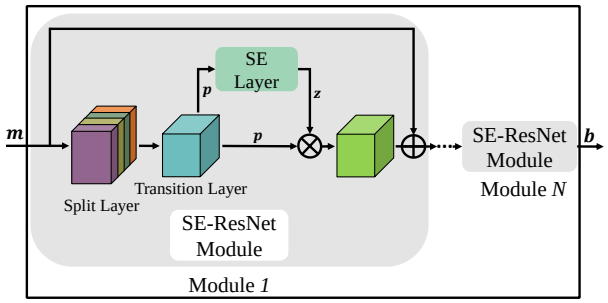


Fig. 3: The proposed system architecture for the speech semantic communication system.



(a) Attention-based semantic encoder.

(b) Attention-based semantic decoder.

Fig. 4: The proposed *semantic encoder* and *semantic decoder* based on SE-ResNet.

a valid 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 *semantic encoder/decoder* and the reasonable training time. Moreover, noise $\boldsymbol{W}$ in the channel layer 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).

2) Testing Stage: Based on the trained networks $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\cdot)$, $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\cdot)$, $\mathbf{R}_{\boldsymbol{\chi}}^{\mathcal{C}}(\cdot)$, and $\mathbf{R}_{\boldsymbol{\delta}}^{\mathcal{S}}(\cdot)$ 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.

The model is trained under a certain fading channel and a fixed SNR value, however, it is impractical to retrain the model for each possible channel condition and load all these models to the transmitter and the receiver. By comparing the testing results for models trained under various channel conditions, we adopt one of these models as the robust model, which could achieve good performance when coping with different channel environments. Accordingly, the robust model is employed to test the performance under various fading channels and different SNR values during the testing stage, as shown in Algorithm 2.

Algorithm 1 Training algorithm of the proposed DeepSC-S

Initialization: initialize parameters $\boldsymbol{\theta}^{\mathcal{T}(0)}$ and $\boldsymbol{\theta}^{\mathcal{R}(0)}$, $i = 0$.

1: **Input:** Speech sample sequences $\boldsymbol{S}$ from speech dataset $\mathfrak{S}$, fading channel $\boldsymbol{H}$, noise $\boldsymbol{W}$ generated under a fixed SNR value.
2: Framing $\boldsymbol{S}$ into $\boldsymbol{m}$ with trainable size.
3: **while** Stop criterion is not meet **do**
4: $\quad\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\boldsymbol{m}) \rightarrow \boldsymbol{b}$.
5: $\quad\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\boldsymbol{b}) \rightarrow \boldsymbol{X}$.
6: $\quad$Transmit $\boldsymbol{X}$ over physical channel and receive $\boldsymbol{Y}$ via
7: $\quad$(2).
8: $\quad\mathbf{R}_{\boldsymbol{\chi}}^{\mathcal{C}}(\boldsymbol{Y}) \rightarrow \widehat{\boldsymbol{b}}$.
9: $\quad\mathbf{R}_{\boldsymbol{\delta}}^{\mathcal{S}}(\widehat{\boldsymbol{b}}) \rightarrow \widehat{\boldsymbol{m}}$.
10: $\quad$Deframing $\widehat{\boldsymbol{m}}$ into $\widehat{\boldsymbol{S}}$.
11: $\quad$Compute loss $\mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}}, \boldsymbol{\theta}^{\mathcal{R}})$ via (4).
12: $\quad$Update trainable parameters simultaneously via SGD:

$$\boldsymbol{\theta}^{\mathcal{T}(i+1)} \leftarrow \boldsymbol{\theta}^{\mathcal{T}(i)} - \eta \nabla_{\boldsymbol{\theta}^{\mathcal{T}(i)}} \mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}}, \boldsymbol{\theta}^{\mathcal{R}}) \quad (8)$$

$$\boldsymbol{\theta}^{\mathcal{R}(i+1)} \leftarrow \boldsymbol{\theta}^{\mathcal{R}(i)} - \eta \nabla_{\boldsymbol{\theta}^{\mathcal{R}(i)}} \mathcal{L}_{MSE}(\boldsymbol{\theta}^{\mathcal{T}}, \boldsymbol{\theta}^{\mathcal{R}}) \quad (9)$$

13: $\quad i \leftarrow i + 1$.
14: **end while**
15: **Output:** Trained networks $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\cdot)$, $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\cdot)$, $\mathbf{R}_{\boldsymbol{\chi}}^{\mathcal{C}}(\cdot)$, and $\mathbf{R}_{\boldsymbol{\delta}}^{\mathcal{S}}(\cdot)$.

V. Experiment and Numerical Results

In this section, we compare the performance of the proposed DeepSC-S, the traditional communication system, and the system with an extra feature coding for speech transmission under the additive white Gaussian noise (AWGN), Rayleigh, and 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 environments, it is tested over telephone systems and multimedia transmission systems.

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 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 transmission systems, the input samples are up-sampled to 44.1KHz. Note that the number of speech samples in different

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

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 values.
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**
    Generate Gaussian noise $\boldsymbol{W}$ under the SNR value.
    $\mathbf{T}_{\boldsymbol{\alpha}}^{\mathcal{S}}(\boldsymbol{m}) \rightarrow \boldsymbol{b}$.
    $\mathbf{T}_{\boldsymbol{\beta}}^{\mathcal{C}}(\boldsymbol{b}) \rightarrow \boldsymbol{X}$.
    Transmit $\boldsymbol{X}$ over physical channel and receive $\boldsymbol{Y}$ via (2).
    $\mathbf{R}_{\boldsymbol{\chi}}^{\mathcal{C}}(\boldsymbol{Y}) \rightarrow \widehat{\boldsymbol{b}}$.
    $\mathbf{R}_{\boldsymbol{\delta}}^{\mathcal{S}}(\widehat{\boldsymbol{b}}) \rightarrow \widehat{\boldsymbol{m}}$.
    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.

.wav is inconsistent. In the simulation, we fix $W = 16,384$, and each sample sequence in $\boldsymbol{S}$ consists of frames $F = 128$ with the frame length $L = 128$.

A. Neural Network Setting and Benchmarks

In the proposed DeepSC-S, the number of SE-ResNet modules in the *semantic encoder/decoder* is 6. For each SE-ResNet module, the number of blocks in the *split* layer is 2, which is achieved by 2 CNN modules with 16 *filters* in each module, and the *transition* layer is implemented by a CNN module with 32 *filters*. Moreover, a single CNN module is utilized in the *channel encoder/decoder*, which contains 8 *filters*. The learning rate is 0.001. The parameters settings of the proposed DeepSC-S in telephone systems are summarized in Table I. For performance comparison, we provide the following three benchmarks.

1) ***Benchmark 1****:* The first benchmark includes the typical source and channel coding techniques. 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 [36]. Moreover, 16-bits PCM is adopted in our work for speech transmission in multimedia transmission systems with $2^{16} = 65,536$ quantization levels. The 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 adopted [37], in which the coding rate is 1/3, the block length is 512, and the number of decoding iterations is 5. 64-QAM is adopted to make the number of transmitted symbols in the traditional communication systems the same as that in DeepSC-S. The details are summarized in Table II.

2) ***Benchmark 2****:* The second benchmark combines feature learning with the traditional channel encoding, named semi-traditional system, as shown in Fig. 5. At the training stage, the *feature encoder* takes the speech samples, $\boldsymbol{s}$, as the inputs and its 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.

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

	Layer Name	Filters	Activation
Transmitter	6×SE-ResNet	6×64	ReLU
	CNN layer	8	None
Receiver	CNN layer	8	ReLU
	6×SE-ResNet	6×64	ReLU
	Last layer (CNN)	1	None

TABLE II: Parameters settings of the traditional communication systems.

	Telephone Systems	Multimedia Systems
Sample rate	8KHz	44.1KHz
Samples 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

For the E2E 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 transmitter and the traditional receiver are same as the settings as shown in Table II. 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 testing stage, the system could yield efficient transmission and mitigate the channel effects. The 2D CNN is adopted in the *feature encoder* and the *feature decoder*. The learning rate is 0.001. The parameters settings are summarized in Table III.

3) ***Benchmark 3****:* In order to emphasize the gain from the attention mechanism, we consider a CNN-based semantic communication system without attention mechanism. The learning rate is 0.001. The parameters settings of the CNN-based system are similar to that of the proposed DeepSC-S as shown in Table I. The only difference is to replace the SE-ResNet modules in the transmitter/receiver for implementing the *semantic encoder/decoder* with the CNN modules. Each CNN module contains 32 *filters*.

B. Complexity Analysis

The complexity of the semi-traditional system is higher than that of the traditional one due to the introduction of the *feature encoder/decoder*. Moreover, due to the complexity of NN training, the CNN-based system and the developed DeepSC-S require higher computational cost than the traditional approaches, which can be measured by the floating

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

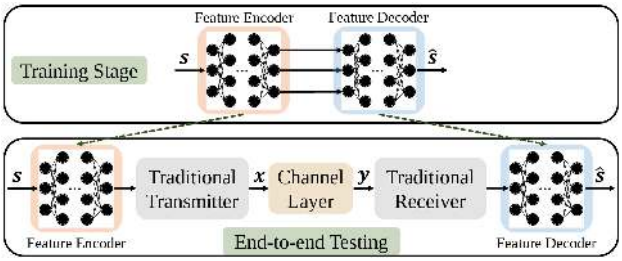


Fig. 5: The benchmark model by combing a feature encoder with the transmission systems.

TABLE III: Parameters settings of the benchmark 2.

	Layer Name	Filters	Activation
Feature Encoder	4×CNN modules	4×32	ReLU
	CNN module	8	None
Feature Decoder	CNN module	8	ReLU
	4×CNN modules	4×32	ReLU
	CNN module	1	None

point operations (FLOPs). For a single 2D CNN module, the required FLOPs can be expressed as [38]:

$$\mathrm{FLOPs_{CNN}} = 2 \times G \times H \times \left(C_{in} \times K^2 + 1\right) \times C_{out}, \quad (10)$$

where G and H are the width and height of the input feature map of the CNN, respectively. K^2 is the kernel size. C_{in} and C_{out} are the number of channels[1] of the input and output feature maps, respectively.

In the semi-traditional system, the number of *filters* of the 5 cascaded CNN modules at the *feature encoder* are 32, 32, 32, 32, 8, respectively. The input size of the *feature encoder* is $128 \times 128 \times 1$, i.e., $G = 128$ and $H = 128$. $K = 5$. C_{in} of the 5 CNN modules are 1, 32, 32, 32, 32, respectively, and C_{out} of the 5 CNN modules are 32, 32, 32, 32, 8, respectively. According to (10), the FLOPs required by the *feature encoder* can be calculated as

$$\begin{aligned}\mathrm{FLOPs}^{\mathcal{T}}_{\mathrm{Semi}} &= 2 \times 128 \times 128 \times (1 \times 5 \times 5 + 1) \times 32 \\ &\quad + 6 \times 128 \times 128 \times (32 \times 5 \times 5 + 1) \times 32 \\ &\quad + 2 \times 128 \times 128 \times (32 \times 5 \times 5 + 1) \times 8 \\ &= 2.75 \times 10^9.\end{aligned} \quad (11)$$

At the receiver, the input size of the *feature decoder* is $128 \times 128 \times 8$ and the number of *filters* of the 6 CNN modules are 8, 32, 32, 32, 32, 1, respectively. Then the FLOPs required by the *feature decoder* are $\mathrm{FLOPs}^{\mathcal{R}}_{\mathrm{Semi}} = 2.81 \times 10^9$. Therefore, the total FLOPs required by the *feature encoder* and the *feature decoder* are 5.56×10^9, i.e., the semi-traditional system requires 5.56×10^9 FLOPs more than the traditional system.

In the CNN-based system, the number of *filters* of the 7 CNN modules at the transmitter are 32, 32, 32, 32, 32, 32, 8, respectively. The input size of the CNN-based system is $128 \times 128 \times 1$, i.e., $G = 128$ and $H = 128$. $K = 5$. C_{in} of the 7 CNN modules are 1, 32, 32, 32, 32, 32, 32, respectively, and C_{out} of the 7 CNN modules are 32, 32, 32, 32, 32, 32, 8, respectively.

[1]Here, the channel refers to the parameter of CNN, not the wireless channels.

According to (10), the FLOPs required by the transmitter are $\mathrm{FLOPs}^{\mathcal{T}}_{\mathrm{CNN}} = 4.44 \times 10^9$. Similarly, the number of *filters* of the 8 CNN modules at the receiver are 8, 32, 32, 32, 32, 32, 32, 1, respectively. Then the FLOPs required by the receiver are $\mathrm{FLOPs}^{\mathcal{R}}_{\mathrm{CNN}} = 4.49 \times 10^9$. Therefore, the total FLOPs required by the CNN-based system are 8.93×10^9.

Furthermore, in the developed DeepSC-S, a single SE-ResNet module consists of 2 CNN modules with 16 *filters* and kernel size is 5×5, as well as 1 CNN module with 32 *filters* and kernel size is 1×1. Given the input size of a single SE-ResNet is $128 \times 128 \times 32$, so $G = 128$ and $H = 128$. C_{in} of the 3 CNN modules are 32, 32, 32, respectively, and C_{out} of the 3 CNN modules are 16, 16, 32, respectively. Then the FLOPs required by this SE-ResNet module are $\mathrm{FLOPs_{SE}} = 8.75 \times 10^8$. Accordingly, the total FLOPs required by DeepSC-S are 9.36×10^9, including 4.65×10^9 FLOPs at the transmitter and 4.71×10^9 FLOPs at the receiver, which achieves a 4.82% increase over the CNN-based system.

C. Experiments over Telephone Systems

In this experiment, we investigate a robust system to work under various channel conditions by training DeepSC-S under the fixed channel condition and then testing it under different fading channels. Note that during the training stage, the Gaussian noise contained in the three training channels is generated under a fixed SNR value, 8 dB, because we found that 8 dB is the most suitable value after comparing the models trained under different SNR values.

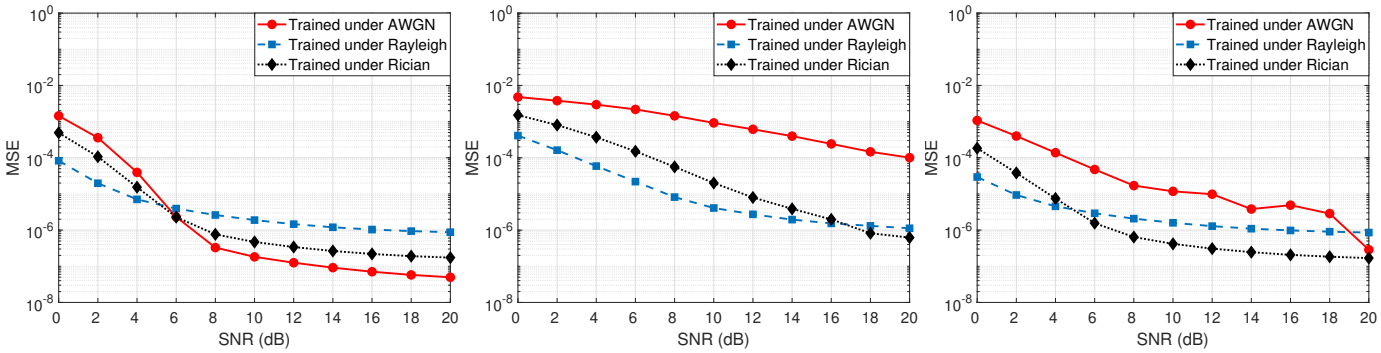
According to Fig. 6, when SNR is lower than 8 dB, DeepSC-S trained under the AWGN channels has higher MSE loss than the models trained under the Rayleigh channels and the Rician channels. As shown in Fig. 6 (a), in terms of the MSE loss tested under the AWGN channels, DeepSC-S trained under the AWGN channels outperforms the model trained under the Rayleigh channels and the Rician channels when SNR is higher than around 6 dB. Besides, according to Fig. 6 (b), DeepSC-S trained under the AWGN channels performs quite poor in terms of MSE loss when testing under the Rayleigh channels. Furthermore, Fig. 6 (c) 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 adopted as a robust model that is capable of coping with various channel environments.

1) SDR and PESQ Results: Based on the roust model, i.e., DeepSC-S trained under the Rician channels with $\mathrm{SNR} = 8$ dB, the relationship between the MSE loss values and the number of epochs is shown in Fig. 7. We can observe that the MSE loss reaches convergence after about 400 epochs.

Fig. 8 tests the SDR performance between the three benchmarks and the proposed DeepSC-S under the AWGN, Rayleigh, and Rician channels. From the figure, the semi-traditional system yields higher SDR score than the traditional one under all tested channel environments while its performance is unreliable when SNR is low. Besides, the CNN-based system and DeepSC-S obtain higher SDR score than

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications



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

Fig. 6: MSE loss tested under (a) AWGN channels, (b) Rayleigh channels, (c) Rician channels with the models trained under various channels.

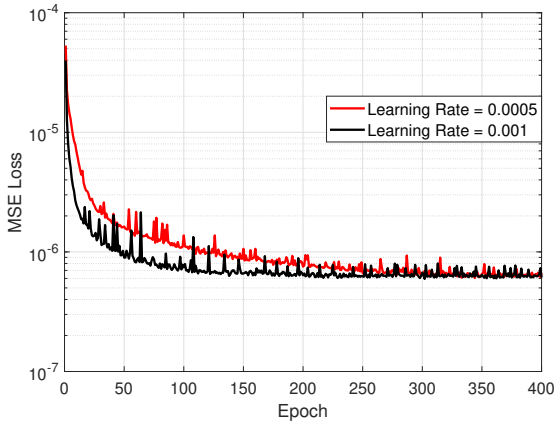


Fig. 7: The training MSE loss versus epoch under the Rician channels with $\mathrm{SNR} = 8$ dB.

the other two systems under the Rayleigh channels and the Rician channels, as well as the AWGN channels over most SNR regions. In addition, DeepSC-S performs steadily when coping with different fading 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 DeepSC-S. Particularly, DeepSC-S achieves 10.92% average ascent over the CNN-based system in terms of the SDR performance.

The PESQ score comparison is shown in Fig. 9. From the figure, the CNN-based system and DeepSC-S outperform the other two systems 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. Particularly, DeepSC-S achieves 7.34% average increase over the CNN-basd system in terms of the PESQ performance. Based on the simulated results, the proposed DeepSC-S is able to yield better speech transmission for telephone systems under complicated communication scenarios than the traditional approaches, especially in the low SNR regime.

D. Experiments over Multimedia Transmission Systems

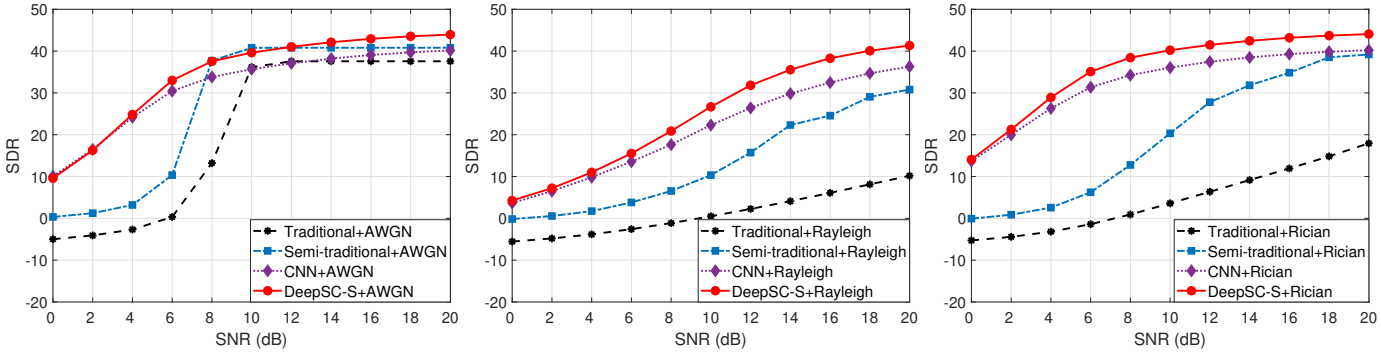
In this part, we present the SDR and PESQ performance comparison. The NN parameters settings of the CNN-based system and DeepSC-S are similar to that in the telephone communications experiment, but the number of *filters* of the CNN layer in both systems is 16. Note that the SDR and PESQ results are also tested under the robust model, i.e., DeepSC-S trained under the Rician channels with $\mathrm{SNR} = 8$ dB.

Fig. 10 depicts the SDR performance comparison for multimedia communications between DeepSC-S and the three benchmarks under the AWGN channels and the Rician channels. Under the AWGN channels, the SDR score of the traditional communication system shows a sharp increase when SNR is higher than 8 dB. Moreover, 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 as 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. 11. From the figure, DeepSC-S outperforms the three benchmarks 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 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.

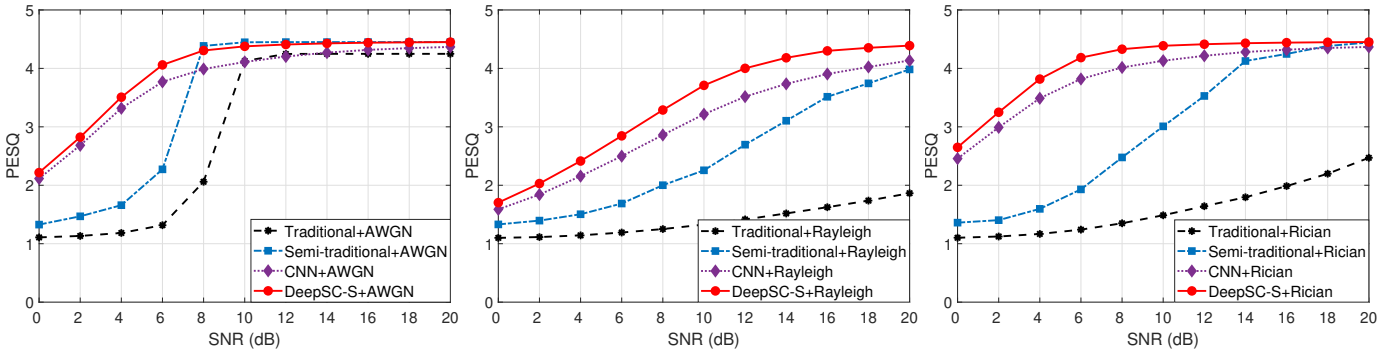
0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications



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

Fig. 8: SDR score versus SNR for speech-based telephone communications with the traditional system, the semi-traditional system, the CNN-based system, and DeepSC-S under (a) AWGN channels, (b) Rayleigh channels, (c) Rician channels.



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

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

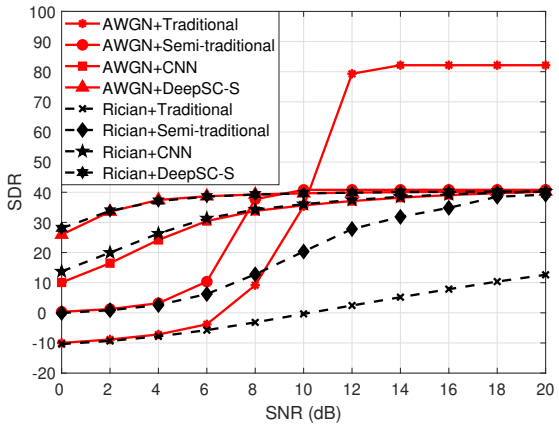


Fig. 10: SDR score versus SNR for speech-based multimedia communications with the traditional system, the semi-traditional system, the CNN-based system, and DeepSC-S under AWGN channels and Rician channels.

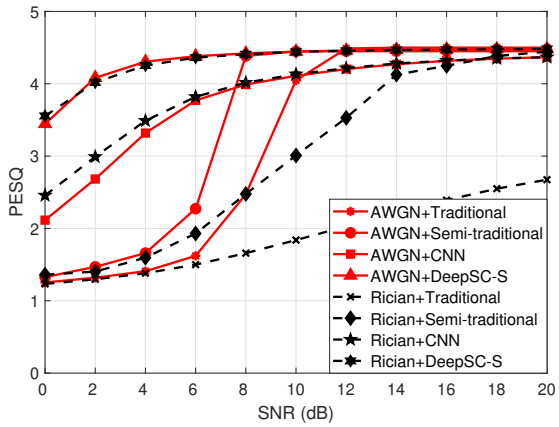


Fig. 11: PESQ score versus SNR for speech-based multimedia communications with the traditional system, the semi-traditional system, the CNN-based system, and DeepSC-S under AWGN channels and Rician channels.

VI. Conclusion

In this article, we investigate a DL-enabled semantic communication system for speech transmission, named DeepSC-S, to improve the transmission efficiency by transmitting the semantic information only. Particularly, we jointly design the *semantic 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 a squeeze-and-excitation network is utilized to improve the recovery accuracy by minimizing the mean-square error of speech signals. Moreover, in order to enable DeepSC-S to

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

work well over various physical channels, a DeepSC-S model with strong robustness to channel variations is developed. The adaptability of the proposed DeepSC-S is verified under the telephone systems and the multimedia transmission systems. Simulation results demonstrate that DeepSC-S outperforms the various benchmarks in the low SNR regime. Hence, the proposed DeepSC-S is a promising candidate for speech semantic communication systems.

REFERENCES

[1] Z. Weng, Z. Qin, and G. Y. Li, "Semantic communications for speech signals," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, Canada, Jun. 2021.

[2] 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.

[3] Z. Qin, G. Y. Li, and H. Ye, "Federated learning and wireless communications," *arXiv preprint arXiv:2005.05265*, May. 2020.

[4] 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.

[5] 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.

[6] N. Samuel, T. Diskin, and A. Wiesel, "Deep MIMO detection," in *Proc. IEEE 18th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1–5.

[7] 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.

[8] 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.

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

[10] R. Carnap and Y. Bar-Hillel, "An outline of a theory of semantic information," Res. Lab. Electron., Massachusetts Inst. Technol., Cambridge, MA, USA, RLE Tech. Rep. 247, Oct. 1952.

[11] A. Maatouk, M. Assaad, and A. Ephremides, "The age of incorrect information: An enabler of semantics-empowered communication," *arXiv preprint arXiv:2012.13214*, Dec. 2020.

[12] 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.

[13] 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.

[14] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 787–802, Dec. 2018.

[15] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, pp. 1–1, Apr. 2021.

[16] 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.

[17] E. Bourtsoulatze, D. B. 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.

[18] 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, May. 2020.

[19] 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.

[20] C.-H. Lee, J.-W. Lin, P.-H. Chen, and Y.-C. Chang, "Deep learning-constructed joint transmission-recognition for internet of things," *IEEE Access*, vol. 7, pp. 76 547–76 561, Jun. 2019.

[21] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Joint device-edge inference over wireless links with pruning," in *Proc. IEEE 21st Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Atlanta, GA, USA, May. 2020, pp. 1–5.

[22] 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, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[23] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[24] 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.

[25] A. Felix, S. Cammerer, S. Dörner, J. Hoydis, and S. Ten Brink, "OFDM-autoencoder for end-to-end learning of communications systems," in *Proc. IEEE 19th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Kalamata, Greece, Jun. 2018, pp. 1–5.

[26] T. J. O'Shea, T. Erpek, and T. C. Clancy, "Physical layer deep learning of encodings for the MIMO fading channel," in *Proc. IEEE 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Monticello, IL, USA, Oct. 2017, pp. 76–80.

[27] H. He, C.-K. 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.

[28] 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.

[29] H. Ye, L. Liang, G. Y. Li, and B.-H. F. Juang, "Deep learning-based end-to-end wireless communication systems with conditional GANs as unknown channels," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3133–3143, May. 2020.

[30] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, 0ct. 2010.

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

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

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

[35] *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, Feb. 2001.

[36] R. V. 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.

[37] Y. Wu and B. D. Woerner, "The influence of quantization and fixed point arithmetic upon the ber performance of turbo codes," in *Proc. IEEE 49th Veh. Technol. Conf. (VTC)*, Houston, TX, USA, May. 1999, pp. 1683–1687.

[38] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, Nov. 2016.

Zhenzi Weng (Graduate Student Member, IEEE) received the B.S. degree from the Ningbo University, Ningbo, China, in 2017, and the M.Sc. degree from the University of Southampton, Southampton, U.K., in 2018. He is currently pursuing the Ph.D. degree with the Queen Mary University of London, U.K., his research interests include semantic communication, channel coding, and machine learning.

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2021.3087240, IEEE Journal on Selected Areas in Communications

Zhijin Qin (Member, IEEE) received the B.S. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2012, and the Ph.D. degree in electronic engineering from the Queen Mary University of London (QMUL), London, U.K., in 2016.

She was a Post-Doctoral Research Associate with Imperial College London from 2016 to 2017 and then, a Lecturer with Lancaster University from 2017 to 2018. Since 2018, she has been a Lecturer with the School of Electronic Engineering and Computer Science, QMUL. Her research interests include semantic communications, deep learning and compressive sensing for wireless signal processing. She was a recipient of the 2017 IEEE GLOBECOM Best Paper Award and the 2018 IEEE Signal Processing Society Young Author Best Paper Award. She serves as an Associate Editor for the IEEE Transactions on Communications, IEEE Communications Letters, and the IEEE Transactions on Cognitive Communications and Networking.

0733-8716 (c) 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Queen Mary University of London. Downloaded on June 11,2021 at 15:59:32 UTC from IEEE Xplore. Restrictions apply.