frontiers in Communications and Networks

ORIGINAL RESEARCH
published: 10 September 2021
doi: 10.3389/frcmn.2021.734402

Federated Learning for Audio Semantic Communication

*Haonan Tong[1], Zhaohui Yang[2], Sihua Wang[1], Ye Hu[3], Omid Semiari[4], Walid Saad[3] and Changchuan Yin[1]**

[1]Beijing University of Posts and Telecommunications (BUPT), Beijing, China, [2]University College London, London, United Kingdom, [3]Virginia Tech, Blacksburg, VA, United States, [4]University of Colorado Colorado Springs, Colorado Springs, CO, United States

In this paper, the problem of audio semantic communication over wireless networks is investigated. In the considered model, wireless edge devices transmit large-sized audio data to a server using semantic communication techniques. The techniques allow devices to only transmit audio semantic information that captures the contextual features of audio signals. To extract the semantic information from audio signals, a wave to vector (wav2vec) architecture based autoencoder is proposed, which consists of convolutional neural networks (CNNs). The proposed autoencoder enables high-accuracy audio transmission with small amounts of data. To further improve the accuracy of semantic information extraction, federated learning (FL) is implemented over multiple devices and a server. Simulation results show that the proposed algorithm can converge effectively and can reduce the mean squared error (MSE) of audio transmission by nearly 100 times, compared to a traditional coding scheme.

Keywords: federated learning, audio communication, semantic communication, autoencoder, wireless network

OPEN ACCESS

Edited by:
Yuanming Shi,
ShanghaiTech University, China

Reviewed by:
Ahmed Imteaj,
Florida International University, United States
Jiawen Kang,
Nanyang Technological University, Singapore

****Correspondence:***
Changchuan Yin
ccyin@bupt.edu.cn

Specialty section:
This article was submitted to Data Science for Communications, a section of the journal Frontiers in Communications and Networks

Received: *01 July 2021*
Accepted: *12 August 2021*
Published: *10 September 2021*

Citation:
Tong H, Yang Z, Wang S, Hu Y, Semiari O, Saad W and Yin C (2021) Federated Learning for Audio Semantic Communication. Front. Comms. Net 2:734402. doi: 10.3389/frcmn.2021.734402

1 INTRODUCTION

Future wireless networks require high data rate and massive connection for emerging applications such as the Internet of Things (IoT) (Saad et al., 2020; Lee et al., 2017; Hu et al., 2021; Al-Garadi et al., 2020; Huang et al., 2021). In particular, in human-computer interaction scenarios, humans may simultaneously control multiple IoT devices using speech, thus making audio communication pervasive in wireless local area network such as smart home. However, due to bandwidth constrains, the wireless network in smart home may not be able to support a broad and prolonged wireless audio communication. This, in turn, motivates the development of semantic communication techniques that allow devices to only transmit semantic information. Semantic communication aims at minimizing the difference between the meanings of the transmitted messages and that of the recovered messages, rather than the recovered symbols. The advantage of such an approach is that semantic communication transmits less amounts of data than traditional communication techniques. However, despite recent interest in semantic communications (Guler et al., 2018; Shi et al., 2020; Xie et al., 2020; Uysal et al., 2021; Xie and Qin, 2021), there is still a lack of reliable encoder and decoder models for audio semantic communication (ASC).

Existing works in Shannon, (1948), Bao et al. (2011), Guler et al. (2018), Shi et al. (2020), Uysal et al. (2021), Xie et al. (2020), Xie and Qin (2021) studied the important problems related to semantic communications. In Shannon (1948), the authors pointed out that semantic communication should consider higher-level information such as content or semantic-related information rather than relying only on data-oriented metrics such as data rate or bit error probability. To efficiently transmit information, the work in Bao et al. (2011) investigated a model-based approach for semantic data

compression and showed that classical source and channel coding theorems have semantic counterparts. Furthermore, the authors in Guler et al. (2018) proposed Bayesian game theory to design the transmission policies for transceivers and minimize the end-to-end average semantic metric while capturing the expected error between the meanings of intended and recovered messages. Besides, the authors in Shi et al. (2020) proposed a semantic-aware network architecture to reduce the required communication bandwidth and significantly improve the communication efficiency. In Uysal et al. (2021), the authors defined a semantic based network system to reduce the data traffic and the energy consumption, hence increasing the wireless devices that can be supported. The work in Xie et al. (2020) proposed a deep learning (DL) based text semantic communication system to reduce wireless traffic load. Meanwhile, in Xie and Qin (2021), the authors developed a new distributed text semantic communication system for IoT devices and they showed that nearly 20 times compression ratio can be achieved without any performance degradation. However, most of these existing works (Shannon, 1948; Bao et al., 2011; Shi et al., 2020; Guler et al., 2018; Uysal et al., 2021; Xie et al., 2020; Xie and Qin, 2021) that focused on the use of semantic communication for text data processing did not consider how to extract the meaning out of the audio data. Here, we note that audio data is completely different from text data since audio signals have a very high temporal resolution, at least 16,000 samples per second (Jurafsky and Martin, 2009).

The prior art in Jurafsky and Martin (2009), Schneider et al. (2019), Amodei et al. (2016), Oord et al. (2016) studied the problem of audio feature extraction. In Jurafsky and Martin (2009), the authors adopted the so-called Mel-frequency cepstral coefficients (MFCC) features to represent the characteristics of audio signals. However, MFCC features are extracted only in a frequency domain, which lacks the contextual relation mining of audio sequence data. Recently, the works in Schneider et al. (2019), Amodei et al. (2016), Oord et al. (2016) used DL based natural language processing (NLP) models to extract audio semantic features. In particular, the authors in (Schneider et al., 2019) proposed a wave to vector (wav2vec) architecture to effectively extract semantic information. The authors in Amodei et al. (2016) proposed an end-to-end model that recognizes various language speeches. In Oord et al. (2016), the authors proposed a speech generator which can generate speech audio signals with different styles using wave data. However, the works in Schneider et al. (2019), Amodei et al. (2016), Oord et al. (2016) did not account for the impact of the channel noise on the transmitted data. Meanwhile, the work in Oord et al. (2016) did not proposed any method to generate the audio signals from the transmitted semantic information.

The use of federated learning (FL) in edge networks was studied in Bonawitz et al. (2019), Tran et al. (2019), Chen et al. (2021a), Chen et al. (2020), Yang K. et al. (2020), Imteaj et al. (2021), Li et al. (2020), Imteaj and Amini (2019), Chen et al., (2021b), Yang Z. et al. (2020), Kang et al. (2019). In Bonawitz et al. (2019), Tran et al. (2019), the authors introduced FL method to generate a global model through collaboratively learning from multiple edge devices, thus learning a distributed algorithm without sharing datasets. The work in Chen et al. (2021c) proposed an FL framework in wireless networks and jointly considered wireless resource allocation and user selection while optimizing FL learning performance. To accelerate the convergence of FL, the authors in Chen et al. (2020) proposed a probabilistic user selection scheme to enhance the efficiency of model aggregation, thus improving convergence speed and the FL training loss. Besides, the authors in Yang K. et al. (2020) introduced over-the-air computation for fast global model aggregation which is realized using superposition property of a wireless multiple-access channel. To explore the applications of FL, the works in Imteaj et al. (2021), Li et al. (2020), Imteaj and Amini (2019), Chen et al. (2021a) provided comprehensive summaries on FL deployed on IoT devices. Besides, the work in Yang Z. et al. (2020) proposed an energy-efficient scheme to minimize the FL energy consumption and complete time, where closed-form solutions of wireless resource allocation are derived. In Kang et al. (2019), the authors proposed efficient incentive mechanisms for FL to improve the learning security and accuracy, which used blockchain based reputation with contract theory. However, most of the above works (Bonawitz et al., 2019; Tran et al., 2019; Chen et al., 2021b; Chen et al., 2020; Yang K. et al., 2020; Imteaj et al., 2021; Li et al., 2020; Imteaj and Amini, 2019; Chen et al., 2021c; Yang Z. et al., 2020; Kang et al., 2019) studied the prediction models which ignored the impact of FL on the performance of semantic communication.

The main contribution of this paper is a novel semantic communication model for audio communication, which is trained via federated learning (FL). Our key contributions include:

- We develop a realistic implementation of an ASC system in which wireless devices transmit large audio command data to a server. For the considered system, the bandwidth for audio data transmission is limited and, thus, semantic information is extracted and transmitted to overcome this limitation. To further improve the accuracy of semantic information extraction, the semantic extraction model must learn from multiple devices. Hence, FL is introduced to train the model with reducing the communication overhead of sharing training data. We formulate this audio communication problem as a signal recovery problem whose goal is to minimize the mean squared error (MSE) between the recovered audio signals and the source audio signals.
- To solve this problem, we propose a wav2vec based autoencoder that uses flexible convolutional neural networks (CNNs) to extract semantic information from source audio signals. The autoencoder consists of an encoder and a decoder. The encoder perceives and encodes temporal features of audio signals into semantic information, which is transmitted over an imperfect wireless channel with noise. Then, the decoder decodes the received semantic information and recovers the audio signals while alleviating channel noise. In this way, the proposed autoencoder transmits less data while jointly designing the source coding and channel coding in the autoencoder.

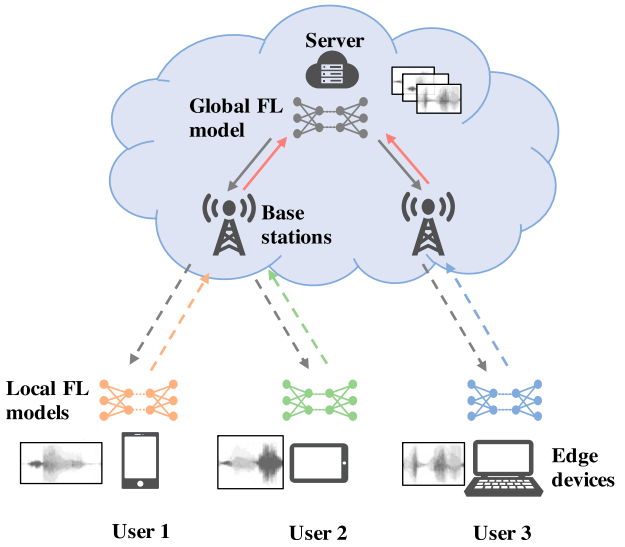

FIGURE 1 | The architecture of an FL based ASC system over wireless networks.

- To improve the accuracy of semantic information extraction, FL is implemented to collaboratively train the autoencoder over multiple devices and the server. In each FL training period, each local model is first trained with the audio data from the local device. Then, the parameters of the local models are transmitted to the server. Finally, the server aggregates the collected local models into a global model and broadcasts the global model to all the devices participated in the FL. Thus, the proposed autoencoder can integrate more audio features from multiple users and, hence, improve the accuracy of semantic information extraction.
- We perform fundamental analysis on the noise immunity and convergence of the proposed autoencoder. We theoretically show that the number of semantic features, time domain downsampling rate, and FL training method can significantly influence performance of the autoencoder.

Simulation results show that the proposed algorithm can effectively converge and reduce the MSE between the recovered and the source audio signals by nearly 100 times, compared to a traditional coding scheme. To our best knowledge, this is the first work that studies the ASC model and uses FL to improve model performance, while avoiding the need for sharing training data.

The rest of this paper is organized as follows. The system model and problem formulation are discussed in *System Model and Problem Formulation*. In *Audio Semantic Encoder and Decoder*, we provide a detailed description of the proposed audio semantic encoder and decoder. The simulation results are presented and analyzed in *Simulation and Performance Analysis*. Finally, conclusions are drawn in *Conclusion*.

2 SYSTEM MODEL AND PROBLEM FORMULATION

We consider a spectrum resource-limited uplink wireless network to deploy an ASC system, which consists of U edge devices, B base stations (BSs), and one server. Each edge device will transmit large audio packets to the server via the closest BS, as shown in **Figure 1**. Due to the limited spectrum, audio semantic information must be extracted for data transmission, thus reducing communication overhead and improving the spectrum efficiency. In particular, edge devices must send audio semantic information *via* wireless channels to the BSs, and, then, the semantic information is delivered via optical links to the server for decoding. To extract the audio semantic information with high efficiency and accuracy, we assume that the edge devices and the server cooperatively train an ASC model using FL. The ASC model consists of an ASC encoder and an ASC decoder, as shown in **Figure 2**. In particular, the ASC encoder is deployed on each edge device to extract audio semantic information while the ASC decoder is deployed on the server to recover audio signals. The objective of the ASC model is to recover the audio signals as accurate as possible. We assume that the connections between BSs and the server use optical links and have sufficient spectrum resource to support accurate transmission. We mainly consider the transmission impairments from the wireless channel between the edge devices and BSs.

To enhance noise immunity, the ASC model must be trained using the received semantic information while taking into account the wireless channel impairments. Hence, the BSs are set to reliably send back the received semantic information to each device, which only occurs during the short-term training stage. Since the extraction of semantic information determines the accuracy of ASC, we consider the architecture design of the ASC model for audio communications.

2.1 ASC Encoder

The ASC encoder is used to encode the input audio data and to extract the semantic information from the raw audio data. We define $\boldsymbol{a} = [a_1, a_2, \ldots, a_T]$ as the raw audio data vector where each element a_t is the audio data in sample t with T being the number of samples. Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]$ be the semantic information vector to be transmitted where x_n is element n in the vector. The ASC encoder extracts $\boldsymbol{x}$ from $\boldsymbol{a}$ by

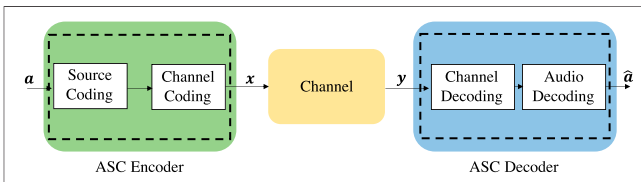

FIGURE 2 | The architecture of audio semantic communication (ASC).

using a neural network (NN) model parameterized by $\boldsymbol{\theta}$, thus, the relationship between $\boldsymbol{a}$ and $\boldsymbol{x}$ can be given by:

$$\boldsymbol{x} = \boldsymbol{T}_{\boldsymbol{\theta}}(\boldsymbol{a}), \tag{1}$$

where $\boldsymbol{T}_{\boldsymbol{\theta}}(\cdot)$ indicates the function of the ASC encoder.

2.2 Wireless Channel

When transmitted over a wireless channel, semantic information will experience channel fading and noise. We assume that the audio transmission uses a single wireless link and, hence, the transmitted signal will be given by:

$$\boldsymbol{y} = h \cdot \boldsymbol{x} + \boldsymbol{\sigma}, \tag{2}$$

where $\boldsymbol{y}$ is the received semantic information at the decoder with transmission impairments, h is the channel coefficient, and $\boldsymbol{\sigma} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ is a Gaussian channel noise at the receiver with variance σ^2. $\boldsymbol{I}$ is the identity matrix.

2.3 ASC Decoder

The ASC decoder is used to recover the audio data $\boldsymbol{a}$ from the received semantic information $\boldsymbol{y}$ and to alleviate transmission impairments. The functions of the decoder and the encoder are generally reciprocal. Let $\hat{\boldsymbol{a}}$ be the decoded audio data and $\boldsymbol{\varphi}$ be the parameters of the NN model in the ASC decoder. Then the relationship between $\hat{\boldsymbol{a}}$ and $\boldsymbol{y}$ can be given by:

$$\hat{\boldsymbol{a}} = \boldsymbol{R}_{\boldsymbol{\varphi}}(\boldsymbol{y}), \tag{3}$$

where $\boldsymbol{R}_{\boldsymbol{\varphi}}(\cdot)$ indicates the function of the ASC decoder.

2.4 ASC Objective

The objective of the ASC system is to recover the audio signals as accurate as possible. Since ASC system transmits semantic information, the use of bit error rate (BER) as a metric is not suitable to assess ASC. Hence, we use the mean squared error (MSE) to evaluate the quality of ASC at the semantic level. The ASC system objective function can be formulated to minimize the MSE between $\boldsymbol{a}$ and $\hat{\boldsymbol{a}}$, as follows:

$$\min_{\boldsymbol{\theta},\boldsymbol{\varphi}} \mathcal{L}_{\mathrm{MSE}}(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{a}, \hat{\boldsymbol{a}}) = \min_{\boldsymbol{\theta},\boldsymbol{\varphi}} \frac{1}{T} \sum_{t=1}^{T} (a_t - \hat{a}_t)^2, \tag{4}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are the parameters of the ASC encoder and ASC decoder, respectively. Here, we assume that the architectures of $\boldsymbol{T}_{\boldsymbol{\theta}}$ and $\boldsymbol{R}_{\boldsymbol{\varphi}}$ are stay fixed and we only update the weights of NNs when solving problem **Equation 4**. Hence, it is necessary to properly design the architecture of the ASC encoder and the ASC decoder. To this end, we introduce an autoencoder to extract audio semantic information.

3 AUDIO SEMANTIC ENCODER AND DECODER

To solve problem (**Eq. 4**), we first propose a wav2vec architecture based autoencoder to efficiently extract audio information. Then, to further improve the accuracy of semantic information extraction, the autoencoder is trained with FL over multiple devices and the server. Thus, the proposed autoencoder can learn semantic information extraction from the audio information of diverse users.

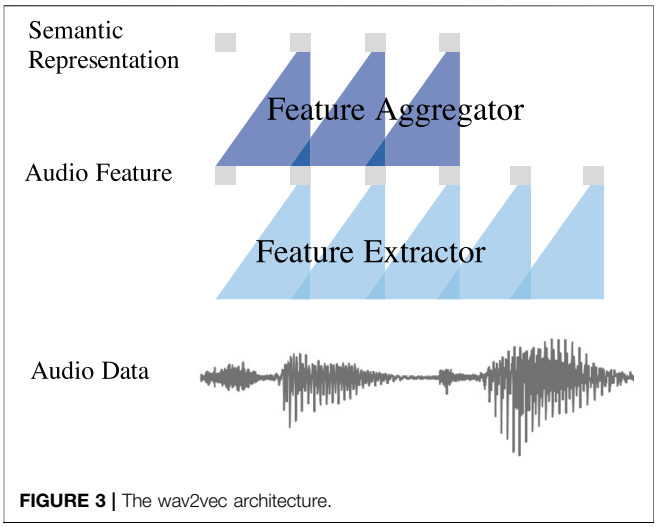

FIGURE 3 | The wav2vec architecture.

3.1 Wav2vec Architecture Based Autoencoder

In the proposed architecture, as shown in **Figure 2**, the ASC system can be interpreted as an autoencoder (O'Shea and Hoydis, 2017; Goodfellow et al., 2016; Lu et al., 2020; Dörner et al., 2018), This autoencoder is trained to recover the input signals at the output end using compressed data features. Since the data must pass through each layer of the autoencoder, the autoencoder must find a robust representation of the input data at each layer (Lu et al., 2020). In particular, NN models are used to build each layer in the autoencoder. Since CNNs are particularly good at extracting features and can be parallel deployed over time on multiple devices, we prefer to use CNNs instead of other NNs such as recurrent neural networks (Shewalkar et al., 2019; Hori et al., 2018; Graves et al., 2013). Next, we introduce our CNN-based wav2vec architecture for semantic information extraction.

To extract the semantic information, we use a wav2vec model as the audio semantic encoder. A simplification of our wav2vec architecture is shown in **Figure 3**. From **Figure 3**, we see that, the wav2vec architecture uses two cascaded CNNs, called feature extractor and feature aggregator (Schneider et al., 2019), to extract audio semantic information. Given the raw audio vector, the extractor refines rough audio features and the aggregator combines the rough audio features into a higher-level latent variable that contains semantic relations among contextual audio features (Schneider et al., 2019).

According to the wav2vec architecture, we design an audio semantic decoder, whose network architecture is symmetrical to the original wav2vec model (Schneider et al., 2019). Combining together an audio semantic encoder and the corresponding semantic decoder, we propose a wav2vec based autoencoder as shown in **Figure 4**. In the

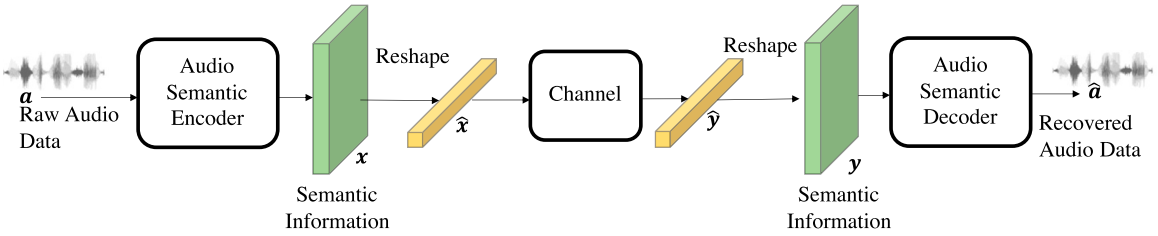

FIGURE 4 | Data shape in the proposed autoencoder over ASC system.

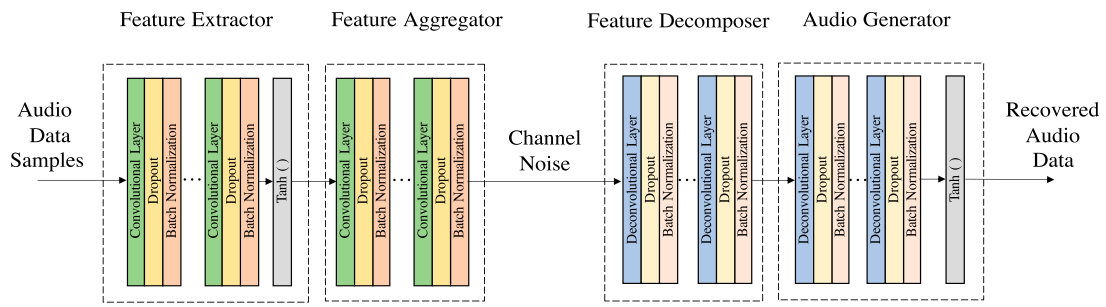

FIGURE 5 | The architecture of the proposed autoencoder.

autoencoder, the audio semantic encoder and the decoder extracts the semantic information and recovers audio signals from the semantic information, respectively. Each single encoder or decoder implements the function of joint source coding and channel coding. Considering the transmission impairments, the semantic information is designed to accurately capture the time domain contextual relations of the audio signals, so as to resist channel fading and noise interference.

Figure 5 shows the NN layers of the proposed autoencoder. According to **Figure 5**, we observe that, given the raw audio signals $\boldsymbol{a}$, the audio semantic encoder is used to extract the semantic vector $\boldsymbol{x}$. In the proposed audio semantic encoder, the data first passes through a feature extractor then a feature aggregator. The feature extractor and the aggregator consist of L_{ext} and L_{agg} convolution blocks, respectively. In particular, each convolution block consists of 1) a convolution layer, 2) a dropout layer, and 3) a batch normalization layer, defined as follows:

- Convolutional Layer: In CNNs, a convolutional layer is used to extract the spatial correlation of the input data with 1-D convolution between the input data $\boldsymbol{Z}^{l-1}$ and the kernel matrix. Mathematically, given the input $\boldsymbol{Z}^{l-1} \in \mathbb{R}^{\lambda_{l-1} \times M^{l-1}}$, the output of the convolutional layer l is $\boldsymbol{Z}^l = [\boldsymbol{z}^{l,1}, \ldots, \boldsymbol{z}^{l,m}, \ldots, \boldsymbol{z}^{l,M^l}], m = 1, \ldots, M^l$, where $\boldsymbol{z}^{l,m} \in \mathbb{R}^{\lambda_l \times 1}$ is the feature map m of convolutional layer l with M^l being the number of output features. Hence, the input $\boldsymbol{Z}^0$ of convolutional layer 1 is the raw audio data or the output of the last NN module. The output of feature map $\boldsymbol{z}^{l,m}$ in each convolutional layer l is given by:

$$\boldsymbol{z}^{l,m} = f\left(\sum_{k=1}^{M^{l-1}} \boldsymbol{z}^{l-1,k} \otimes \boldsymbol{W}_c^{l,m} + \boldsymbol{b}_c^{l,m}\right), \tag{5}$$

where $f(x) = x$ is the linear activation function, M^{l-1} is the number of feature maps in the last convolutional layer l—1, $\otimes$ denotes 1-D convolution operation, and $\boldsymbol{W}_c^{l,m} \in \mathbb{R}^{s_k \times 1}$ and $\boldsymbol{b}_c^{l,m}$ are convolution kernels and bias vector of feature map m in convolutional layer l, respectively, with s_k being the kernel size. Let the convolution stride be s_c, the padding size be p, and the size of feature map λ_l satisfies $\lambda_l = \left\lfloor \frac{\lambda_{l-1} + 2p - s_k}{s_c} \right\rfloor + 1$.

- Dropout Layer: The input $\boldsymbol{Z}^l$ of a dropout layer l is the output of convolutional layer l. In the training stage, the dropout layer randomly abandons the effect of each neuron with a probability called dropout rate, and, in the inference stage, the dropout layer counts on the effects of all neurons. The dropout layer is used as a regularization approach to avoid overfitting problem.
- Batch Normalization Layer: A batch normalization layer normalizes the values of activated neurons to avoid gradient vanishing. We define α_i as the value of the activated neuron i in convolution block l. The normalized value $\hat{\alpha}$ of the neuron is given by $\hat{\alpha}_i = \frac{\alpha_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$, where $\mu_{\mathcal{B}} = \frac{1}{\lambda_l}\sum_{i=1}^{\lambda_l} \alpha_i$, $\sigma_{\mathcal{B}}^2 = \frac{1}{\lambda_l}\sum_{i=1}^{\lambda_l} (\alpha_i - \mu_{\mathcal{B}})^2$, and ϵ is a positive constant.

Since the amplitude of an audio signal is limited, $\tanh(\cdot)$ is introduced as the activation function of the output layer in the feature extractor (Oord et al., 2016), where $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. To shape the transmitted semantic information with an adequate amplitude, the last

layer of the feature aggregator is set as batch normalization layer without activation function (Dörner et al., 2018).

In the proposed audio semantic decoder, as shown in **Figure 5**, the received semantic information first passes through a feature decomposer then an audio generator. Different from the encoder, a deconvolution operation is introduced to build the feature decomposer and audio generator which consist of L_{de} and L_{gen} deconvolution blocks, respectively. Correspondingly, each deconvolution block consists of 1) one deconvolution layer, 2) one dropout layer, and 3) one batch normalization layer. Mathematically, the processes of the dropout layer and the batch normalization layer are similar to those in the convolution blocks, except for the deconvolution layer.

In the deconvolution layer, the feature matrix is first uniformly filled with zeros in each column. Given the filled input matrix $\tilde{\boldsymbol{Z}}^{l-1} = [\tilde{\boldsymbol{z}}^{l-1,1}, \ldots, \tilde{\boldsymbol{z}}^{l-1,m}, \ldots, \tilde{\boldsymbol{z}}^{l-1,M^{l-1}}] \in \mathbb{R}^{\tilde{\lambda}_{l-1} \times M^{l-1}}$, the output of a deconvolution layer l is $\boldsymbol{Z}^l = [\boldsymbol{z}^{l,1}, \ldots, \boldsymbol{z}^{l,m}, \ldots, \boldsymbol{z}^{l,M^l}] \in \mathbb{R}^{\lambda_l \times M^l}, m = 1, \ldots, M^l$, where $\tilde{\boldsymbol{z}}^{l,m}$ is the filled feature map m and $\tilde{\lambda}_{l-1}$ is the filled feature map size. The output of feature map $\boldsymbol{z}^{l,m}$ in each convolutional layer l is given by:

$$\boldsymbol{z}^{l,m} = f\left(\sum_{k=1}^{M^{l-1}} \tilde{\boldsymbol{z}}^{l-1,k} \otimes \boldsymbol{W}_d^{l,m} + \boldsymbol{b}_d^{l,m}\right), \tag{6}$$

where M^{l-1} is the number of features of layer l–1, and $\boldsymbol{W}_d^{l,m}$ and $\boldsymbol{b}_d^{l,m}$ are deconvolution kernels and bias vector in deconvolutional layer l, respectively. In deconvolution layer, the filled feature map size $\tilde{\lambda}_{l-1}$ satisfies $\tilde{\lambda}_l = s_c(\lambda_{l-1} - 1) - 2p + 2s_k - 1$ and the size of feature map λ_l satisfies $\lambda_l = \tilde{\lambda}_{l-1} - s_k + 1 = s_c(\lambda_{l-1} - 1) - 2p + s_k$, where p is the padding size of layer l. Note that, to appropriately recover the audio signals, the output layer of the audio generator is set as tanh(·) function.

To amplify the inference error and avoid gradient vanishing, we introduces the normalized root mean squared error (NRMSE) for the autoencoder. Then the objective of the autoencoder is given by:

$$\min_{\boldsymbol{\theta},\boldsymbol{\varphi}} \ \mathcal{L}_{\text{NRMSE}}(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{a}, \hat{\boldsymbol{a}}) = \min_{\boldsymbol{\theta},\boldsymbol{\varphi}} \ \frac{\sum_{t=1}^{T} (a_t - \hat{a}_t)^2}{\sum_{t=1}^{T} a_t^2}. \tag{7}$$

Algorithm 1 Local model training algorithm of the autoencoder.

1: **Initialize:** Randomly initialize parameters $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\varphi}^{(0)}$, initialize training epoch $i = 0$.
2: **Input:** Batches of audio data $\boldsymbol{a}$
3: **while** model does not converge **or** $i <$ max epoch **do**:
4: $\boldsymbol{T}_{\boldsymbol{\theta}}(\boldsymbol{a}) \rightarrow \boldsymbol{x}$.
5: Transmit $\boldsymbol{x}$ over wireless channel and receive $\boldsymbol{y}$.
6: $\boldsymbol{R}_{\varphi}(\boldsymbol{y}) \rightarrow \widehat{\boldsymbol{a}}$.
7: Calculate $\mathcal{L}_{NRMSE}(\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{a}, \widehat{\boldsymbol{a}})$ according to (9).
8: Update encoder's and decoder's parameters simultaneously with SGD:

$$\begin{aligned} \boldsymbol{\theta}^{(i+1)} &\leftarrow \boldsymbol{\theta}^{(i)} - \eta \nabla_{\boldsymbol{\theta}^{(i)}} \mathcal{L}_{\text{NRMSE}}\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\varphi}^{(i)}\right) \\ \boldsymbol{\varphi}^{(i+1)} &\leftarrow \boldsymbol{\varphi}^{(i)} - \eta \nabla_{\boldsymbol{\varphi}^{(0)}} \mathcal{L}_{\text{NRMSE}}\left(\boldsymbol{\theta}^{(i)}, \boldsymbol{\varphi}^{(i)}\right) \end{aligned} \tag{8}$$

9: $i \leftarrow i + 1$
10: **end while**
11: **Output:** $\boldsymbol{T}_{\boldsymbol{\theta}}(\cdot)$ and $\boldsymbol{R}_{\varphi}(\cdot)$ combined as $\boldsymbol{w}$.

3.2 FL Training Method

Next, our goal is to minimize the errors between the recovered audio signals and the source audio signals using FL training method. In FL, the server and the devices collaboratively learn the proposed autoencoder by sharing the model parameters (Liu et al., 2020; Chen et al., 2021a; Chen et al., 2021b; Yang et al., 2021). We define $\boldsymbol{w} = (\boldsymbol{\theta}, \boldsymbol{\varphi})$ as the total parameter of the proposed autoencoder, which includes both the encoder and decoder. The server generates a global model $\boldsymbol{w}^{\boldsymbol{g}}$ and each device i locally trains a local autoencoder model $\boldsymbol{w}_i$ which shares the same architecture as $\boldsymbol{w}^{\boldsymbol{g}}$, as shown in **Figure 1**. The global model periodically aggregates local models from U devices that participate in FL and broadcasts the aggregated global model back to the devices. Then the aggregated global model can be given by $\boldsymbol{w}^{\text{g}} = \frac{1}{U}\sum_{i=1}^{U} \boldsymbol{w}_i$. We use $\boldsymbol{A}_i$ to capture the audio dataset of local model i. According to problem **Eq. 7**, the objective of FL training method is given by:

$$\min_{\boldsymbol{w}^{\text{g}}} \ \sum_{i=1}^{U} \mathcal{L}_{\text{NRMSE}}\left(\boldsymbol{w}_i, \boldsymbol{A}_i, \widehat{\boldsymbol{A}}_i\right). \tag{9}$$

Algorithm 2 FL training algorithm of the global model (Imteaj et al., 2021).

1: **Initialization:** Initialize the model architecture, the local models and the global model are initialized with random parameters. Initialize model aggregation step $\tau = 0$.
2: **while** model does not converge **or** $\tau <$ max aggregation step **do**:
3: Each local model updates $\boldsymbol{w}_1^{(\tau)}, \boldsymbol{w}_2^{(\tau)} \ldots \boldsymbol{w}_U^{(\tau)}$ through training from local dataset according to Algorithm 1.
4: Transmit local trained models $\boldsymbol{w}_1^{(\tau)}, \boldsymbol{w}_2^{(\tau)} \ldots \boldsymbol{w}_U^{(\tau)}$ to the server.
5: Update the global model: $\boldsymbol{w}^{\text{g}(\tau)} = \frac{1}{U}\sum_{i=1}^{U} \boldsymbol{w}_i^{(\tau)}$
6: Dispatch local models: $\boldsymbol{w}_1^{(\tau+1)}, \boldsymbol{w}_2^{(\tau+1)} \ldots \boldsymbol{w}_U^{(\tau+1)} \leftarrow \boldsymbol{w}^{\text{g}(\tau)}$
7: $\tau \leftarrow \tau + 1$
8: **end while**
9: **Output:** Global model $\boldsymbol{w}^{\text{g}}$.

During the local model training stage, the server first defines the architecture of the autoencoder and broadcasts it to all edge devices to randomly initialize the local models. To keep the coordination between the encoder and the decoder of the proposed autoencoder, we jointly set that the encoder and the decoder update the parameters simultaneously to minimize the loss function **Eq. 9**. Hence, both the encoder and decoder update the parameters with stochastic gradient descent (SGD) once after a batch of data passes through the autoencoder.The training process of each local model can be shown in **Algorithm 1**, where η in (8) is the learning rate. During the training process of the global model, each edge device is set to transmit the parameters of the local models $\boldsymbol{w}_i$ to the server every a fixed number of epochs. Thus, the server periodically collects the transmitted models, aggregates the parameters of the local models, and then broadcasts the updated global model to each device. In the next period, the local models update their parameters through training from local datasets $\boldsymbol{A}_i$, before transmitting $\boldsymbol{w}_i$ to the server, as shown in **Algorithm 1**. The FL algorithm for the global model is summarized in **Algorithm 2**.

TABLE 1 | Simulation parameters.

Module	Setting	Parameter	Value
feature extractor	$L_{ext} = 3$	feature M^l	8,8,8
		kernel size s_k	1,2,4
		stride s_c	1,1,1
		dropout rate	0.5
feature aggregator	$L_{agg} = 4$	feature M^l	8,8,8,8
		kernel size s_k	2,4,8,16
		stride s_c	1,1,1,1
		dropout rate	0.5
feature decomposer	$L_{de} = 4$	feature M^l	8,8,8,8
		kernel size s_k	2,4,8,16
		stride s_c	1,1,1,1
		dropout rate	0.5
audio generator	$L_{gen} = 4$	feature M^l	8,8,8,1
		kernel size s_k	1,2,4,1
		stride s_c	1,1,1,1
		dropout rate	0.5

3.3 Complexity Analysis

The proposed FL algorithm used to solve problem **Eq. 9** is summarized in **Algorithm 2**. The complexity of the proposed algorithm lies in training the proposed autoencoder. The complexity for training the autoencoder is $\mathcal{O}(\sum_{l=1}^{L}\lambda_l^2 {s_k}^2 M^l M^{l-1})$ (Wang et al., 2020), where $L = L_{ext} + L_{agg} + L_{de} + L_{gen}$, with L_{ext}, L_{agg}, L_{de}, L_{gen}, and L being the number of convolution or deconvolution layers in the feature extractor, the feature aggregator, the feature decomposer, the audio generator, and the proposed autoencoder, respectively. Let L_o be the number of model aggregations until the FL global model converges. The complexity of the FL training method is $\mathcal{O}(L_o U \sum_{l=1}^{L}\lambda_l^2 {s_k}^2 M^l M^{l-1})$ (Chen et al., 2020). In consequence, the major complexity of training the autoencoder, which depends on the number of NN layers, the kernel sizes and the numbers of features in each layer, is linear. Meanwhile, since the layers in the autoencoder are finite, the local training is achievable and, hence the edge devices can support the FL training in the considered wireless network. Once the training process is completed, the trained autoencoder can be used for ASC in a long term period.

4 SIMULATION AND PERFORMANCE ANALYSIS

To evaluate the proposed autoencoder, we train the model using a training set from the speech dataset Librispeech (Panayotov et al., 2015), which contains 1,000 h of 16 kHz read English speech. The learning rate η is 10^{-5}. The proposed autoencoder is trained under additive white Gaussian noise (AWGN) channels with a fixed channel coefficient h and a 6dB signal-to-noise-ratio (SNR), and it is tested on 200,000 samples of speech data. The simulation parameters are listed in **Table 1** (Kang et al., 2020). We train the model using FL method with 1 global model and 2 local models of user 1 and user 2, each local model is trained using read speech from a single person, and the FL models are tested with read speech of another user 3. The global model aggregates local models every 10 local training epochs.

For comparison purposes, we simulate a baseline scheme for high-quality audio transmission, which uses 128 kbps pulse code modulation (PCM) with 8 bits quantization levels (Nakano et al., 1982) for source coding, low-density parity-check codes (LDPC) (Gallager, 1962) for channel coding, and 64-QAM (Pfau et al., 2009) for modulation. In this section, for notational convenience, we call the proposed autoencoder for ASC a "semantic method," and we call the baseline scheme a "traditional method". Note that, the autoencoder is trained *via* NRMSE, and tested *via* MSE. This is because that NRMSE induces larger gradient for training the autoencoder and MSE provides more obvious fluctuations for result comparison. To verify the performance of the proposed FL algorithm, we compare two baselines: transfer learning method and local gradient descent FL (Imteaj et al., 2021). In the transfer learning method, the feature aggregator

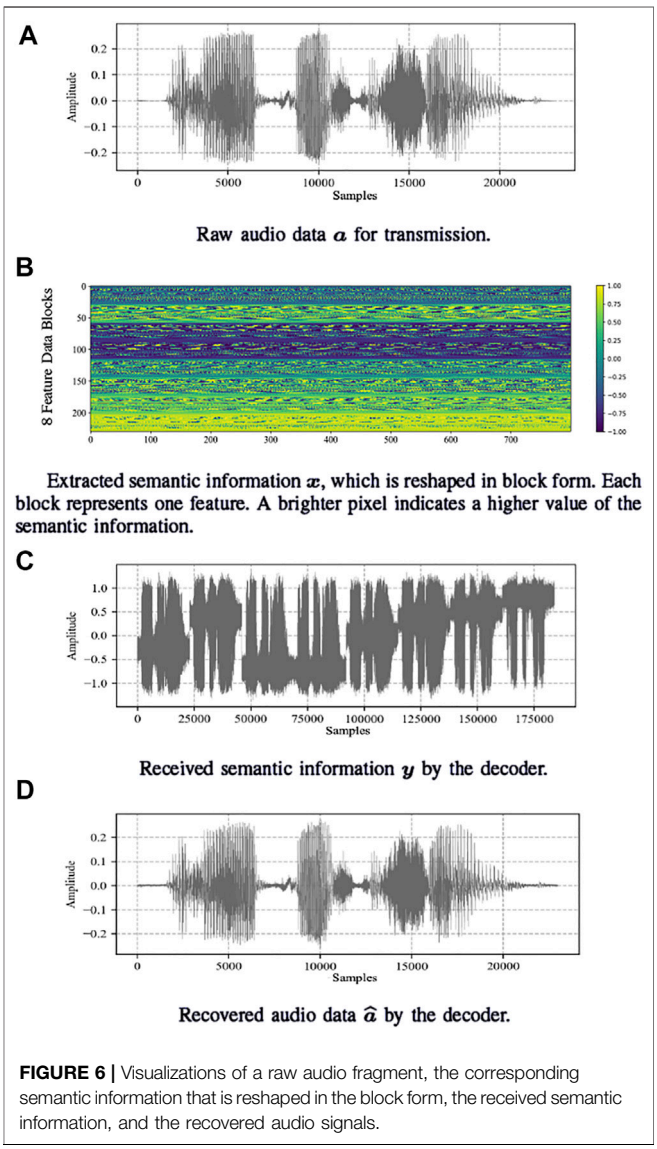

FIGURE 6 | Visualizations of a raw audio fragment, the corresponding semantic information that is reshaped in the block form, the received semantic information, and the recovered audio signals.

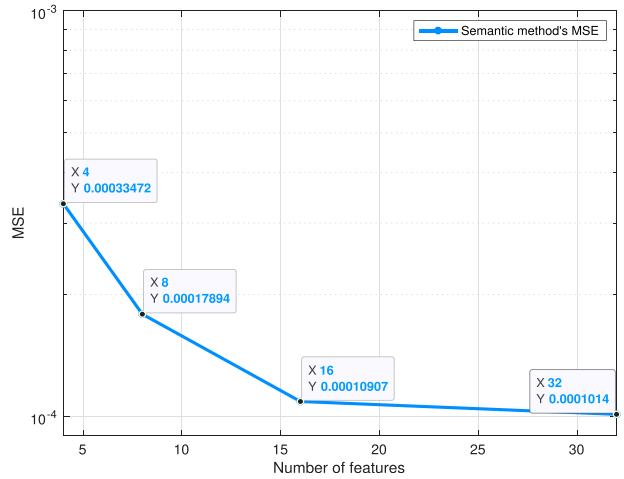

FIGURE 7 | Transmission MSE of a local autoencoder model as the number of features varies, in AWGN channels with a 6dB SNR.

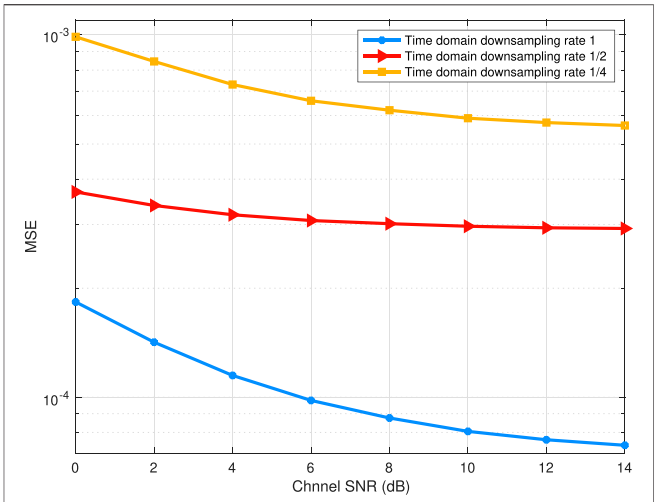

FIGURE 9 | Transmission MSE of the proposed semantic method with different time domain downsampling rates. The number of semantic features is 8.

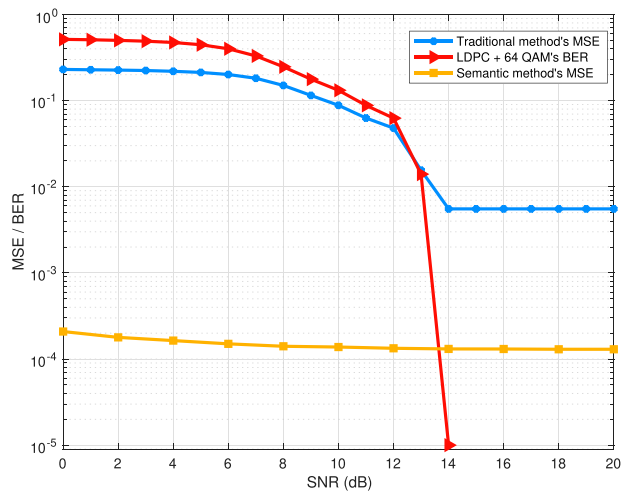

FIGURE 8 | Transmission MSE of a local model using semantic method, BER and transmission MSE of traditional method as SNR varies.

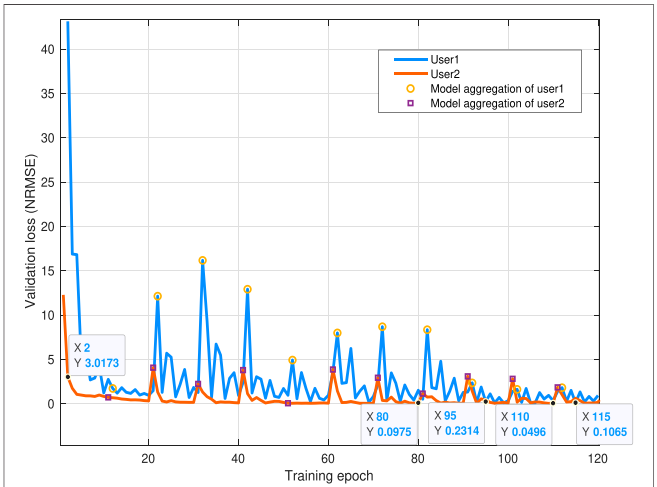

FIGURE 10 | Convergence results of the proposed FL models.

and the decomposer in the autoencoder are first initialized with a pre-trained model, then the autoencoder is trained using local audio data. In the local gradient descent FL, at the start of each iteration, all devices first share an aggregated model, then each device simultaneously computes a fixed number of local gradient descent updates (1,000 steps) in parallel.

Figure 6 shows examples of the raw audio data, the extracted semantic information reshaped in block form, the received semantic information, and the recovered audio data in one local model. From **Figures 6A–C**, we see that, the audio semantic information signals are amplified by the proposed semantic encoder before being transmitted through the channel. From **Figure 6B**, it is also observed that, the extracted eight different blocks of semantic features have correlations. From **Figures 6C,D**, we see that the proposed semantic decoder eliminates the channel noise from the received signals. The elimination of the noise is due to the fact that the semantic decoder relieves the noise using multiple semantic features. **Figure 6** shows that the proposed autoencoder can effectively guarantee the accuracy of ASC.

Figure 7 shows how transmission MSE of a local model using semantic method changes as the number of features varies. From **Figure 7**, we see that, as the number of features increases, the MSE of the proposed semantic method decreases first and, then remains unchanged. This phenomenon is due to the fact that higher dimension features provide better semantic representations thus improving the transmission performance of the semantic

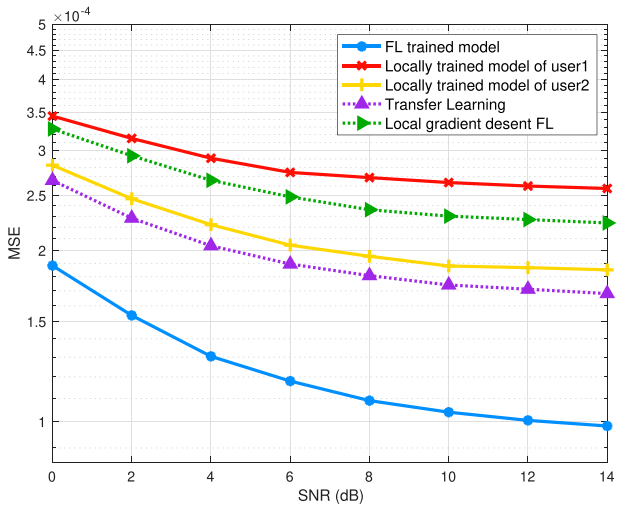

FIGURE 11 | Transmission MSE of FL trained model, locally trained models, transfer learning and local gradient descent FL.

method. From **Figure 7**, we can also see that, when the number of features is larger than 16, the MSE of the semantic method tends to be leveling off. This result is because of the existence of redundant semantic features which provide limited noise immunity for ASC.

In **Figure 8**, we show how the transmission MSE of a local model using the proposed semantic method, BER and MSE of the traditional method change as the channel SNR varies. In this simulation, the semantic method reduces communication overhead by decreasing nearly 1/3 of the transmission data amount compared to the traditional method. From **Figure 8**, we observe that, as the channel SNR increases, the error of communication decreases as expected. From **Figure 8**, we can also see that our semantic method reduces the transmission MSE by nearly 100 times, compared to the traditional method, and the MSE of semantic method varies flatter than that of traditional method. The improvement is due to the fact that the semantic method has a better transmission accuracy and noise immunity performance. From **Figure 8**, we can also see that, the MSE of the traditional method remains unchanged when the SNR is larger than 14 dB. The phenomenon is because, for a lower BER, the accuracy of the traditional coding scheme will reach the coding limit and, hence, the MSE will stay at a quantization error level caused by PCM quantization.

Figure 9 shows how the transmission MSE changes versus various channel SNR, where the semantic method uses different time domain downsampling rates. In this simulation, lower time domain downsampling rates can reduce transmission data amount exponentially and are realized by changing the convolution strides in the feature extractor and feature decomposer. From **Figure 9**, we can see that, a lower time domain downsampling rate leads to more transmission error, which is because of the more loss of semantic information. From **Figure 9**, we can also observe that as the SNR increases, the decreasing speed of the MSE differs among different downsampling rates. The disparity is due to the fact that, the semantic information extracted with different downsampling rates has diverse sensitivities to the SNR. **Figure 9** shows that reducing the time domain sampling rates decreases the communication accuracy. In consequence, **Figure 7** and **Figure 9** demonstrate that, in terms of improving the performance of semantic communication, the complexity of semantic features trades off the data compression rate.

In **Figure 10**, we show how the validation loss changes as the training epoch increases. From **Figure 10**, we observe that, the validation loss initially decreases with fluctuation first and then remains unchanged. The fact that the validation loss remains unchanged demonstrates that the FL algorithm converges. From **Figure 10**, we can see that, when the FL global model is aggregated, the loss of local models increases for several epochs first and, then decreases in a long-term view. The result is due to the difference of the multiple local audio datasets from different users. At the beginning of training, the aggregation of multiple local models will critically change the parameter distribution of the global model. Then, as the training process continues, the global model parameters fit multiple local datasets. Hence the fluctuation caused by FL model aggregation weakens, and the local models of multiple users converge. From **Figure 10**, we can also see that FL model aggregation further decreases the lower bound of loss in each local model. This phenomenon is because that FL training method aggregates audio semantic features from multiple users, thus enhancing model performance compared with local training method.

Figure 11 shows how the transmission MSE of all algorithms changes as the channel SNR varies. From **Figure 11**, we observe that the performance of the proposed model differs among the diverse users due to the various audio characteristics. We can also see that transfer learning can improve the model performance compared to locally training. Besides, local gradient descent FL outperforms part, but not all of the locally trained models. The difference of the baselines is because that transfer learning can further learn audio semantic extraction based on pre-trained model parameters. Whilst local gradient descent FL aggregates the global model with low frequency, where the difference among local models leads to the inefficiency on improving semantic extraction. From **Figure 11**, we can also see that, the proposed FL algorithm outperforms the locally trained models. The superiority is because that the FL trained model aggregates audio characteristics of all users and hence obtaining more robust performance. We can also observe from the dotted lines that the proposed FL training method is superior over transfer learning and local gradient descent FL. The superiority is due to the fact that the proposed FL algorithm aggregates the model in a frequent and synchronous way, which guarantees a more accurate semantic extraction than that of the baselines.

5 CONCLUSION

In this paper, we have developed an FL trained model over an ASC architecture in the wireless network. We have considered avoidance of training data sharing and heavy communication overhead of the large-sized audio transmission between edge devices and the server. To solve this problem, we have proposed a wav2vec based autoencoder to effectively encode, transmit, and decode audio semantic information, rather than traditional bits or symbols, to reduce communication overhead. Then, the autoencoder is trained with FL to improve the accuracy of semantic information extraction. Simulation results have shown that the proposed algorithm can converge effectively and yields significant reduction on transmission error compared to existing coding scheme which uses PCM, LDPC and 64-QAM algorithm.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary files, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

HT: formulate the audio semantic communication system, use the learning approach, and do the simulation results. ZY: help proof reading the whole paper and provide suggestions about the structure of the paper. SW: help checking the simulation results and debug the codes. YH: provide suggestions about the flow chart and write the theoretical analysis. OS: do the literature review and identify the key novelty of this paper. WS: provide the idea about audio semantic communication and guide the writing. CY: polish the language of the whole paper and check all the formulations.

FUNDING

This work was supported in part by Beijing Natural Science Foundation and Municipal Education Committee Joint Funding Project under Grant KZ201911232046, in part by the National Natural Science Foundation of China under Grants 61671086 and 61629101, in part by the 111 Project under Grant B17007, in part by U.S. National Science Foundation (NSF) under Grants CNS-2007635 and CNS-2008646, and in part by BUPT Excellent Ph.D. Students Foundation under Grant CX2021114.

REFERENCES

Al-Garadi, M. A., Mohamed, A., Al-Ali, A. K., Du, X., Ali, I., and Guizani, M. (2020). A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security. *IEEE Commun. Surv. Tutorials* 22 (3), 1646–1685. doi:10.1109/COMST.2020.2988293

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., et al. (2016). "Deep Speech 2: End-To-End Speech Recognition in English and Mandarin," in Proc. of International Conference on Machine Learning (NY, USA: ICML), 173–182.

Bao, J., Basu, P., Dean, M., Partridge, C., Swami, A., Leland, W., et al. (2011). Towards a Theory of Semantic Communication. *Proc. IEEE Netw. Sci. Workshop* 2011, 110–117. doi:10.1109/nsw.2011.6004632

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., et al. (2019). Towards Federated Learning at Scale: System Design. *Arxiv, Vol. abs/1902.01046*. doi:10.1109/TWC.2020.3042530

Chen, M., Gündüz, D., Huang, K., Saad, W., Bennis, M., Feljan, A. V., et al. (2021a). Distributed Learning in Wireless Networks: Recent Progress and Future Challenges. *arXiv:2104.02151*. [Online]. Available: http://arxiv.org/abs/2104.02151.

Chen, M., Shlezinger, N., Poor, H. V., Eldar, Y. C., and Cui, S. (2021b). Communication-efficient Federated Learning. *Proc. Natl. Acad. Sci.* 118 (17), e2024789118. doi:10.1073/pnas.2024789118

Chen, M., Yang, Z., Saad, W., Yin, C., Poor, H. V., and Cui, S. (2021c). A Joint Learning and Communications Framework for Federated Learning over Wireless Networks. *IEEE Trans. Wireless Commun.* 20 (1), 269–283. doi:10.1109/twc.2020.3024629

Chen, M., Vincent Poor, H., Saad, W., and Cui, S. (2020). Convergence Time Optimization for Federated Learning over Wireless Networks. *IEEE Trans. Wireless Commun.* 20 (4), 2457–2471. doi:10.1109/TWC.2020.3042530

Dörner, S., Cammerer, S., Hoydis, J., and Brink, S. T. (2018). Deep Learning Based Communication over the Air. *IEEE J. Sel. Top. Signal. Process.* 12 (1), 132–143. doi:10.1109/jstsp.2017.2784180

Gallager, R. (1962). Low-density Parity-Check Codes. *IEEE Trans. Inform. Theor.* 8 (1), 21–28. doi:10.1109/tit.1962.1057683

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). "Speech Recognition with Deep Recurrent Neural Networks," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (Vancouver, BC, Canada: ICASSP). doi:10.1109/icassp.2013.6638947

Guler, B., Yener, A., and Swami, A. (2018). The Semantic Communication Game. *IEEE Trans. Cogn. Commun. Netw.* 4 (4), 787–802. doi:10.1109/tccn.2018.2872596

Hori, T., Cho, J., and Watanabe, S. (2018). "End-to-end Speech Recognition with Word-Based RNN Language Models," in Proc. IEEE Spoken Language Technology Workshop (Athens, Greece: SLT), 389–396. doi:10.1109/slt.2018.8639693

Hu, Y., Chen, M., Saad, W., Poor, H. V., and Cui, S. (2021). Distributed Multi-Agent Meta Learning for Trajectory Design in Wireless Drone Networks. *IEEE J. Selected Areas Commun.* PP (99), 1. doi:10.1109/jsac.2021.3088689

Huang, C., Yang, Z., Alexandropoulos, G. C., Xiong, K., Wei, L., Yuen, C., et al. (2021). Multi-hop RIS-Empowered Terahertz Communications: A DRL-Based Hybrid Beamforming Design. *IEEE J. Selected Areas Commun.* doi:10.1109/jsac.2021.3071836

Imteaj, A., and Amini, M. H. (2019). "Distributed Sensing Using Smart End-User Devices: Pathway to Federated Learning for Autonomous IoT," in Proc. International Conference on Computational Science and Computational Intelligence (Las Vegas, NV, USA: CSCI).

Imteaj, A., Thakker, U., Wang, S., Li, J., and Amini, M. H. (2021). A Survey on Federated Learning for Resource-Constrained IoT Devices. *IEEE Internet Things J.* doi:10.1109/jiot.2021.3095077

Jurafsky, D., and Martin, J. H. (2009). *Speech and Language Processing*. NJ, USA: Prentice-Hall.

Kang, J., Xiong, Z., Niyato, D., Xie, S., and Zhang, J. (2019). Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to Combining Reputation and Contract Theory. *IEEE Internet Things J.* 6 (6), 10 700–710 714. doi:10.1109/jiot.2019.2940820

Kang, J., Xiong, Z., Niyato, D., Zou, Y., Zhang, Y., and Guizani, M. (2020). Reliable Federated Learning for mobile Networks. *IEEE Wireless Commun.* 27 (2), 72–80. doi:10.1109/mwc.001.1900119

Lee, S. K., Bae, M., and Kim, H. (2017). Future of IoT Networks: A Survey. *Appl. Sci.* 7 (10), 1072. doi:10.3390/app7101072

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. (2020). Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal. Process. Mag.* 37 (3), 50–60. doi:10.1109/msp.2020.2975749

Liu, Y., Yuan, X., Xiong, Z., Kang, J., Wang, X., and Niyato, D. (2020). Federated Learning for 6G Communications: Challenges, Methods, and Future Directions. *China Commun.* 17 (9), 105–118. doi:10.23919/jcc.2020.09.009

Lu, Y., Cheng, P., Chen, Z., Li, Y., Mow, W. H., and Vucetic, B. (2020). Deep Autoencoder Learning for Relay-Assisted Cooperative Communication Systems. *IEEE Trans. Commun.* 68 (9), 5471–5488. doi:10.1109/tcomm.2020.2998538

Nakano, K., Moriwaki, H., Takahashi, T., Akagiri, K., and Morio, M. (1982). A New 8-bit Pcm Audio Recording Technique Using an Extension of the Video Track. *IEEE Trans. Consumer Electron.* CE-28 (3), 241–249. doi:10.1109/tce.1982.353917

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., et al. (2016). Wavenet: A Generative Model for Raw Audio. *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499.

O'Shea, T., and Hoydis, J. (2017). An Introduction to Machine Learning Communications Systems. *ArXiv:1702.00832, Vol. abs/1702.00832*. [Online]. Available: http://arxiv.org/abs/1702.00832.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., (2015). "Librispeech: An Asr Corpus Based on Public Domain Audio Books," in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (Queensland, Australia: ICASSP), 5206–5210.

Pfau, T., Hoffmann, S., and Noe, R. (2009). Hardware-Efficient Coherent Digital Receiver Concept with Feedforward Carrier Recovery for M-QAM Constellations. *J. Lightwave Technol.* 27 (8), 989–999. doi:10.1109/jlt.2008.2010511

Saad, W., Bennis, M., and Chen, M. (2020). A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems. *IEEE Netw.* 34 (3), 134–142. doi:10.1109/mnet.001.1900287

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). Wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv:1904.05862*. [Online]. Available: http://arxiv.org/abs/1904.05862.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell Syst. Tech. J.* 27 (3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shewalkar, A., Nyavanandi, D., and Ludwig, S. A. (2019). Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU. *J. Artif. Intelligence Soft Comput. Res.* 9 (4), 235–245. doi:10.2478/jaiscr-2019-0006

Shi, G., Xiao, Y., Li, Y., and Xie, X. (2020). From Semantic Communication to Semantic-Aware Networking: Model, Architecture, and Open Problems. *arXiv:2012.15405*. [Online]. Available: http://arxiv.org/abs/2012.15405.

Tran, N. H., Bao, W., Zomaya, A., Nguyen, M. N. H., and Hong, C. S. (2019). "Federated Learning over Wireless Networks: Optimization Model Design and Analysis," in Proc. IEEE Conference on Computer Communications (Paris, France: IEEE).

Uysal, E., Kaya, O., Ephremides, A., Gross, J., Codreanu, M., Popovski, P., et al. (2021). Semantic Communications in Networked Systems. *arXiv:2103.05391*. [Online]. Available: http://arxiv.org/abs/2103.05391.

Wang, Y., Chen, M., Yang, Z., Luo, T., and Saad, W. (2020). Deep Learning for Optimal Deployment of UAVs with Visible Light Communications. *IEEE Trans. Wireless Commun.* 19 (11), 7049–7063. doi:10.1109/TWC.2020.3007804

Xie, H., and Qin, Z. (2021). A Lite Distributed Semantic Communication System for Internet of Things. *IEEE J. Select. Areas Commun.* 39 (1), 142–153. doi:10.1109/jsac.2020.3036968

Xie, H., Qin, Z., Li, G. Y., and Juang, B.-H. (2020). Deep Learning Enabled Semantic Communication Systems. *arXiv:2006.10685*. [Online]. Available: http://arxiv.org/abs/2006.10685.

Yang, K., Jiang, T., Shi, Y., and Ding, Z. (2020). Federated Learning via Over-the-air Computation. *IEEE Trans. Wireless Commun.* 19 (3), 2022–2035. doi:10.1109/TWC.2019.2961673

Yang, Z., Chen, M., Saad, W., Hong, C. S., and Shikh-Bahaei, M. (2020). Energy Efficient Federated Learning over Wireless Communication Networks. *IEEE Trans. Wireless Commun.* 20 (3), 1935–1949. doi:10.1109/TWC.2020.3037554

Yang, Z., Chen, M., Wong, K.-K., Poor, H. V., and Cui, S. (2021). Federated Learning for 6G: Applications, Challenges, and Opportunities. *arXiv:2101.01338*. [Online]. Available: http://arxiv.org/abs/2101.01338.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.