ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 1,Jan. 2026,pp. 117－131
DOI:10. 16337/j. 1004-9037. 2026. 01. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于个性化联邦学习和语义通信的语音传输系统

刘月照，郭海燕，王添顺，陈飞飞

（南京邮电大学通信与信息工程学院，南京 210003）

摘　要：面向多用户语音传输场景，本文提出一种使用超网络个性化联邦学习的深度学习语义通信系统（Deep learning based semantic communication system using federated learning based on hypernetworks，DeepSC-FedHN）。边缘服务器采用超网络来衡量每个本地用户语义编码器中各模块的重要性，生成个性化聚合权重矩阵来更新相应模型参数。同时，采用联邦学习（Federated learning，FL）算法聚合模型的信道编解码器和语义解码器部分。实验结果表明，本文提出的DeepSC-FedHN方案总体优于本地训练方案、联邦平均（Federated averaging，FedAvg）方案、联邦近似（Federated proximal，FedProx）方案和采用分层个性化联邦学习的深度学习语义通信系统（Deep learning based semantic communication system using layer-wised personalized federated learning，DeepSC-pFedLA）。

关键词：语音传输；语义通信；联邦学习；超网络

中图分类号：TN929.5　　**文献标志码：**A

引用格式：刘月照，郭海燕，王添顺，等. 基于个性化联邦学习和语义通信的语音传输系统[J]. 数据采集与处理，2026，41(1)：117-131. LIU Yuezhao，GUO Haiyan，WANG Tianshun，et al. Speech transmission system based on personalized federated learning and semantic communication[J]. Journal of Data Acquisition and Processing，2026，41(1)：117-131.

引　　言

近年来，随着深度学习技术的高速发展，语义通信研究受到越来越多的关注，成为目前的研究热点之一[1-2]。与传统通信旨在传输恢复源信号自身数据不同，语义通信关注与接收端任务（如分类、检测、重构等）相关的语义信息的提取和传输，从而有效节约了通信资源[3]。目前，国内外学者已开展面向文本类任务[4-6]、图像类任务[7-9]、视频类任务[10-11]和多模态信号类任务[12-13]等的语义通信研究，结果表明，语义通信与传统通信相比，大幅度提高了通信效率，有效改进了用户的体验质量[14]。

在面向语音类任务的语义通信研究方面，文献[15-18]针对语音识别任务开发了一系列语义通信模型。文献[19-20]针对语音翻译任务设计了不同的语义通信模型。针对语音传输任务，Weng等[21]提出利用挤压和激励（Squeeze-and-excitation，SE）网络捕捉语音信号的基本特征，基于SE-ResNet模块，设计了基于深度学习的语音语义通信系统（Deep learning enabled semantic communication system for speech signals，DeepSC-S）。Xiao等[22]使用联合信源信道编解码（Joint source-channel coding，JSCC）方法设计了命名为深度语音语义传输（Deep speech semantic transmission，DSST）的语音传输系统，在保证语音重构质量前提下减少了模型的信道带宽。文献[23]将语音识别和语音合成作为通信系统的传输任务，提出了基于深度学习的语音传输语义通信系统（Deep learning based semantic communication system for speech transmission，DeepSC-ST），在接收端通过将识别的文本和说话人信息输入到神经网

基金项目：国家自然科学基金(62071242)。
收稿日期：2024-09-29；**修订日期：**2025-02-18

络(Neural network, NN)模块,可以重新生成语音信号。文献[24]利用Transformer-XL捕获长距离依赖的能力,设计了一个用于实时语音传输的语义通信系统。文献[25]提出了一个基于扩散模型的音频语义通信系统,该系统可以在信道条件高度退化的情况下对语音信号进行去噪和修复。

上述语音传输类任务的语义通信研究主要面向单用户场景。然而,在现实环境中,通常多个用户都有语音传输需求。并且,单个用户的本地数据往往并不充分,仅通过单个用户的本地数据训练本地模型,可能会导致模型对于语义信息提取的性能较弱,而多用户采用联邦学习(Federated learning, FL)框架[26]进行协作,可以在保障用户数据隐私的前提下更新得到性能更佳的本地模型,从而更有效地提取各用户的语义信息。

FL作为一种新兴的分布式机器学习框架,其主要思想是基于分布在多个设备上的数据集构建机器学习模型,同时防止数据泄露。将FL引入到多用户语义通信中,可以有效地聚合更新各个用户的本地模型,提高模型的性能。Tong等[27]面向语音传输任务,开发了基于FL的语义通信系统,采用联邦平均(Federated averaging, FedAvg)算法进行模型的聚合更新,显著降低了边缘设备与服务器之间的通信开销。面向图像传输任务,Deng等[28]提出了一种结合卷积神经网络(Convolutional neural network, CNN)和Transfromer的语义通信模型,采用联邦平均FedAvg算法来聚合用户上传的模型,显著提高了模型的性能,并且减少了每个用户的训练时间。Xie等[29]提出了一种基于FL的语义通信框架,用于物联网设备的多任务分布式图像传输。Wei等[30]提出了一种联邦语义学习框架(Federated semantic learning, FedSem),在基于基站的语义信道解码器的协调下,协同训练多个设备的语义信道编码器,可以充分利用分布式数据和计算资源。Xie等[31]面向车牌识别任务,开发了一个自动编码器来执行语义编码,同时采用异步联邦学习算法,以确保训练过程能够容忍传输延迟。文献[32]面向文本传输任务提出了一种语义通信系统模型的FL部署方式,能够使模型有效地学习到用户数据的特征。

在上述采用FL的多用户语义通信研究中,通常采用的是FedAvg算法。该算法根据各设备上的本地数据集量的大小,为每个设备上传的模型参数赋予一个权重,然后对模型参数进行加权平均。然而FedAvg算法没有考虑到不同用户数据集之间存在的非独立同分布(Non independent and identically distributed, Non-IID)特性[33],导致所有用户共享同一个全局模型可能会与各用户的本地数据偏离较大,影响其语义信息的提取性能。个性化联邦学习(Personalized federated learning, pFL)机制通过优化FL模型聚合过程来构建个性化模型,可以在全局模型的基础上为每个用户训练一个定制的模型,以适应自己的数据分布[34-35]。

鉴于上述考虑,本文面向多用户语音传输场景,提出一种使用超网络[36]个性化联邦学习的深度学习语义通信系统(Deep learning based semantic communication system using federated learning based on hypernetworks, DeepSC-FedHN)。考虑到各个用户本地的语音数据存在说话人身份、地区或性别等差异,利用FL框架,采用超网络为每个用户生成一个个性化的聚合权重矩阵,对每个本地基础模型DeepSC-S[21]进行个性化更新。本文构建了一种结合pFL和FedAvg的多用户语音语义传输系统。该系统考虑到各用户语音数据分布的差异性,结合FedAvg和pFL,对各用户的本地模型进行聚合更新,获得更匹配本地数据分布特性的个性化本地模型,提高各用户学习提取语音语义信息的能力。然后,提出了一种基于超网络的改进pFL算法。该算法考虑到语义编码器的不同模块在语义信息的学习和提取的过程中发挥的作用不同,使用超网络为每个用户的语义编码器模块生成个性化的聚合权重矩阵,以产生个性化语义编码器参数。同时,考虑到信道编解码器和语义解码器不参与本地用户数据语义特征的提取,使用FedAvg算法对各用户的信道编解码器和语义解码器进行加权聚合更新。在TIMIT数据集和Edinburgh DataShare语音数据集上的实验结果表明,本文提出的DeepSC-FedHN在客观语音质量评估(Perceptual evaluation of speech quality, PESQ)、信号失真比(Signal-to-distortion ration,

SDR)和短时客观可懂度(Short-time objective intelligibility, STOI)上均优于本地训练策略、基于FedAvg的训练策略和基于FedProx[37]的训练策略。另外,本文所提出的DeepSC-FedHN的模型聚合计算量显著低于采用分层个性化联邦学习的深度学习语义通信系统(Deep learning based semantic communication system using layer-wised personalized federated learning, DeepSC-pFedLA),对未知说话人数据的泛化性也更好。

1 相关工作

1.1 DeepSC-S模型

DeepSC-S模型结构如图1所示。DeepSC-S模型由语义编码器、信道编码器、信道解码器和语义解码器组成。在发送端,语音信号分帧后先通过语义编码器学习到表征语义信息的特征,再经信道编码器将语义特征编码为符号序列,经由无线信道传输。在接收端,接收到的特征先通过信道解码器解码,再经过语义解码器解码得到恢复的语音帧,进而对各帧进行重叠相加,重构语音信号。

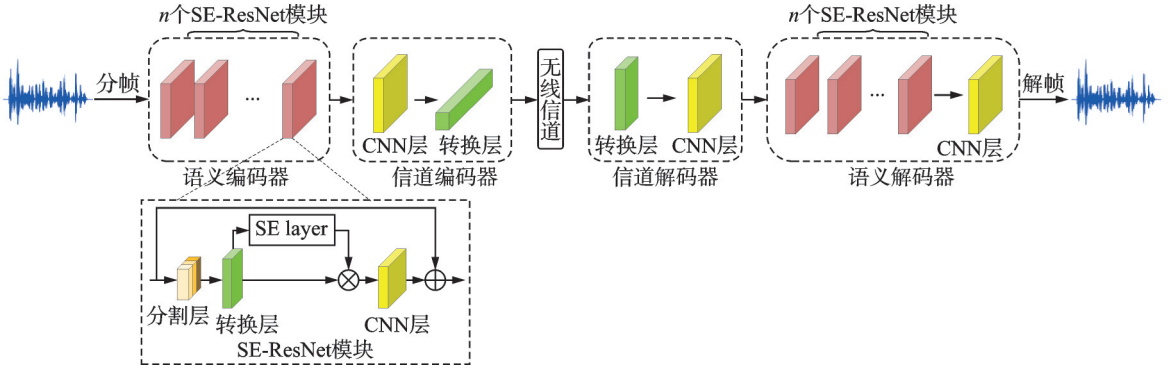


图1 DeepSC-S模型结构

Fig.1 Structure of DeepSC-S model

在DeepSC-S模型中,信道编码器和信道解码器各由1个包含2维卷积模块的CNN层构成,语义编码器和语义解码器由若干个基于注意力机制的SE-ResNet模块组成。SE-ResNet模块包含1个分割层、1个过渡层、1个SE层和1个CNN层,该模块被用于学习和提取语音信号的基本特征。

1.2 超网络

超网络是一种为另一个神经网络生成权重的神经网络,一般由若干个简单的全连接层组成,其输入是可学习的嵌入。超网络可以在训练过程中动态地生成适应当前任务的权重,因此可以用来为每个用户生成个性化本地模型。文献[38]首次将超网络用于pFL,在服务器上学习1个超网络,为每个用户的本地模型直接输出个性化模型。与文献[38]不同,文献[39]使用超网络为各用户本地模型的每一层输出聚合权重,从而得到个性化的本地模型。在文献[38-39]中,基于超网络的pFL方法都是为CNN模型设计的,并没有考虑到更为复杂的模型。考虑到文献[38-39]在处理复杂模型方面的局限性,本文针对语音语义传输模型,提出采用超网络为各用户语义编码器中的各个模块生成不同的聚合权重来更新本地模型参数,这与文献[38]采用超网络直接输出个性化模型、文献[39]采用超网络输出每一层聚合权重的方法不同。

2 系统模型

本文考虑在频谱资源受限的无线网络部署一个面向多用户语音传输的语义通信系统。该系统由N

对用户和1个边缘服务器BS组成，每个用户都采用设备到设备(Device to device, D2D)方式进行无线通信，如图2所示。例如在智能家居或智慧工厂等场景中，各发送端用户向接收端用户传输语音指令等信息。在图2中，本地用户对U_i、U_i'训练一个本地的DeepSC-S模型，然后将训练好的本地模型的语义编码器参数θ_i、信道编码器参数ϕ_i、信道解码器参数χ_i和语义解码器参数φ_i上传到边缘服务器BS中，BS通过模型聚合模块更新全局模型。

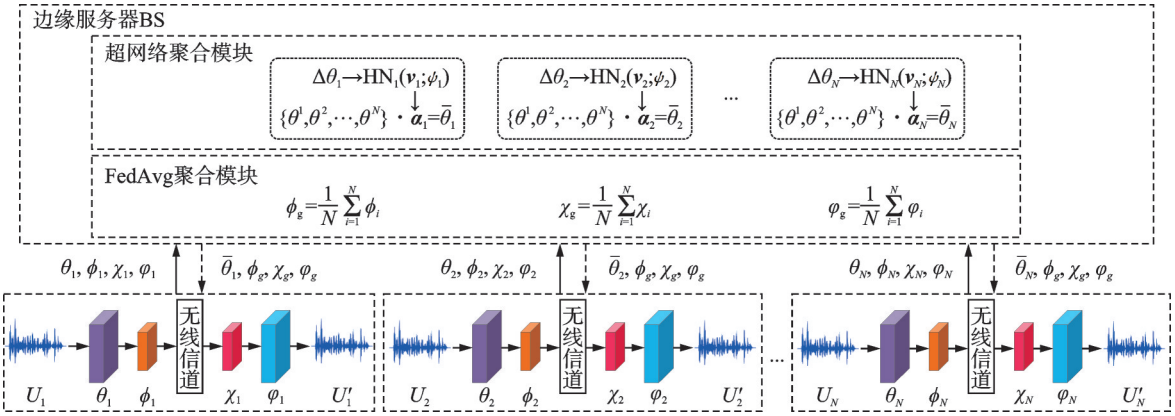


图2 DeepSC-FedHN系统结构图

Fig.2 Architecture diagram of DeepSC-FedHN system

边缘服务器BS的模型聚合模块由超网络聚合模块和FedAvg聚合模块组成。其中，超网络聚合模块由N个超网络$\mathrm{HN}_1(\boldsymbol{\nu}_1;\boldsymbol{\psi}_1)$，$\mathrm{HN}_2(\boldsymbol{\nu}_2;\boldsymbol{\psi}_2)$，…，$\mathrm{HN}_N(\boldsymbol{\nu}_N;\boldsymbol{\psi}_N)$组成。$\mathrm{HN}_i(\boldsymbol{\nu}_i;\boldsymbol{\psi}_i)$为第$i$个用户对$U_i$、$U_i'$生成个性化的聚合权重矩阵$\boldsymbol{\alpha}_i$，然后根据$\boldsymbol{\alpha}_i$生成个性化的语义编码器参数$\bar{\theta}_i$。FedAvg聚合模块采用加权平均的聚合方式对各DeepSC-S模型的其他部分进行聚合更新，得到全局模型ϕ_g、χ_g、φ_g。

一旦全局模型更新完成，边缘服务器BS将$\bar{\theta}_i$、ϕ_g、χ_g、φ_g返回给各用户，用户将本地模型更新为接收到的全局模型，并启动本地模型训练过程，进行下一轮的FL训练。

2.1 用户本地模型训练

在每个FL训练轮次中，每个用户使用本地数据集训练各自的DeepSC-S模型。以第i个用户对U_i、U_i'为例进行说明，这里为简便起见，将参数下标省略。原始语音序列$\boldsymbol{s}$经过语义编码器$S_\theta(\cdot)$和信道编码器$C_\phi(\cdot)$，得到编码后的语义信息序列$\boldsymbol{x}$，即

$$\boldsymbol{x}=C_\phi(S_\theta(\boldsymbol{s})) \tag{1}$$

记信道参数为$\boldsymbol{h}$，则接收端接收信号$\boldsymbol{y}$为

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

式中：$\boldsymbol{n}\sim N(0,\sigma^2 I)$表示均值为0、方差为$\sigma^2$的高斯噪声，“$*$”表示卷积运算。

接收信号$\boldsymbol{y}$经信道解码器$C_\chi(\cdot)$和语义解码器$S_\varphi(\cdot)$后，恢复的解码信号$\hat{\boldsymbol{s}}$为

$$\hat{\boldsymbol{s}}=S_\varphi(C_\chi(\boldsymbol{y})) \tag{3}$$

采用均方误差(Mean-squared error, MSE)损失函数L_{MSE}进行本地训练，L_{MSE}定义为

$$L_{\mathrm{MSE}}(\theta,\phi,\chi,\varphi)=\sum_{j=1}^{B}(\boldsymbol{s}_j-\hat{\boldsymbol{s}}_j)^2 \tag{4}$$

式中B为批次大小。

用户完成本地训练后,将各自的本地模型参数,即$\{\theta_1,\phi_1,\chi_1,\varphi_1\},\{\theta_2,\phi_2,\chi_2,\varphi_2\},\cdots,\{\theta_N,\phi_N,\chi_N,\varphi_N\}$发送至服务器进行模型聚合。

2.2 服务器模型聚合更新

边缘服务器BS接收到各用户的本地模型参数后,通过模型聚合模块对模型参数进行聚合。模型聚合模块由基于FedAvg算法的聚合模块和超网络聚合模块两部分组成。其中,由于信道编解码器和语义解码器不参与语义信息的提取,采用基于FedAvg算法[26]的聚合模块对信道编码器、信道解码器和语义解码器的模型参数进行更新,以得到一个泛化性更高的公共模型。

2.2.1 超网络聚合模块

鉴于各用户数据分布有差异,并且语义编码器中的各SE-ResNet模块在提取语义特征的过程中发挥的作用有所不同[21],本文考虑在服务器上为每个用户定制一个专用的超网络,为这个用户的各SE-ResNet模块生成聚合权重。

如图3所示,超网络由若干个全连接层组成,在最后一层全连接层后紧接着归一化处理,其输入是嵌入向量$\boldsymbol{\nu}_i$,输出权重矩阵$\boldsymbol{\alpha}_i$为

$$\boldsymbol{\alpha}_i=\mathrm{HN}_i(\boldsymbol{\nu}_i;\ \psi_i)=\{\boldsymbol{\alpha}_i^1,\boldsymbol{\alpha}_i^2,\cdots,\boldsymbol{\alpha}_i^n\}=\begin{bmatrix}\alpha_i^{1,1} & \alpha_i^{1,2} & \cdots & \alpha_i^{1,n}\\ \alpha_i^{2,1} & \alpha_i^{2,2} & \cdots & \alpha_i^{2,n}\\ \vdots & \vdots & \ddots & \vdots\\ \alpha_i^{N,1} & \alpha_i^{N,2} & \cdots & \alpha_i^{N,n}\end{bmatrix} \tag{5}$$

式中:ψ_i为面向用户U_i的超网络的参数,$\boldsymbol{\alpha}_i^n$表示第n个SE-ResNet模块的聚合权重向量,$\alpha_i^{j,n}$表示用户U_j第n个SE-ResNet模块的聚合权重。通过在超网络的输出前添加归一化处理,使得式(5)满足$\sum_{j=1}^{N}\alpha_i^{j,n}=1$,即所有用户第$n$个SE-ResNet模块权重的和为1。超网络的输出权重矩阵$\boldsymbol{\alpha}_i$包含了语义编码器中各SE-ResNet模块的聚合权重,通过模型聚合更新的方式,能够有效地加强各个用户彼此间之间的协作。

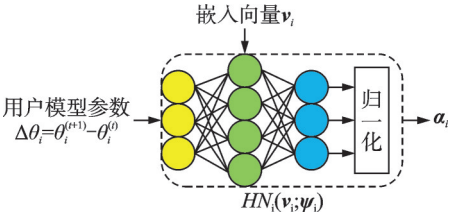


图3 超网络的工作流程

Fig.3 Workflow of hypernetworks

服务器收到各用户完成本地训练的模型参数后,使用超网络聚合模块为每个用户生成个性化的语义编码器模型。具体地,设$\theta_i=\{\theta_i^1,\theta_i^2,\cdots,\theta_i^n\}$为用户$U_i$经过本地训练后的语义编码器中SE-ResNet块的模型参数,其中θ_i^n表示用户U_i的第n个SE-ResNet块的参数。令$\theta^n=\{\theta_1^n,\theta_2^n,\cdots,\theta_N^n\}$表示所有用户第$n$个SE-ResNet块参数的集合,根据超网络$HN_i(\boldsymbol{\nu}_i;\ \psi_i)$的输出$\boldsymbol{\alpha}_i$,用户$U_i$的语义编码器中SE-ResNet块的模型参数更新为

$$\bar{\theta}_i=\{\bar{\theta}_i^1,\bar{\theta}_i^2,\cdots,\bar{\theta}_i^n\}=\{\theta^1,\theta^2,\cdots,\theta^n\}\cdot\boldsymbol{\alpha}_i \tag{6}$$

式中:$\bar{\theta}_i^n=\sum_{j=1}^{N}\theta_j^n\alpha_i^{j,n}$,"$\cdot$"表示点积运算。

2.2.2 模型聚合计算量分析

单个CNN层的参数量可以表示为[40]

$$P=C_{\mathrm{out}}\times(C_{\mathrm{in}}\times K^2+1) \tag{7}$$

式中:C_{out}表示输出通道数,C_{in}表示输入通道数,K表示卷积核大小。由文献[21]可得,单个SE-ResNet块由3个CNN层组成,则单个SE-ResNet块参数量为$3P$。超网络输出权重矩阵$\boldsymbol{\alpha}_i$的大小为$N\times n$,因此得到$\bar{\theta}_i$需要的计算量为$3P\times N\times n$。DeepSC-S模型信道编解码器各由1层CNN层组成,语义解码器由$3n+1$层CNN层组成,因此FedAvg聚合模块的计算量为$P\times(3n+1)\times N$,聚合模块得到$\bar{\theta}_i$、ϕ_g、

χ_g、φ_g，总的模型聚合计算量为$3P\times N\times n+P\times(3n+1)\times N$。

而采用类似文献[39]的方法，令超网络为DeepSC-S模型的所有层生成聚合权重时，模型的总层数为$6n+3$，超网络输出的聚合权重矩阵大小为$N\times(6n+3)$，则得到个性化模型需要的计算量为$P\times N\times(6n+3)^2$。

3 训练过程

用户利用本地的数据对各自的本地模型进行训练，经过E个训练轮次后，将其模型参数上传到云端服务器进行模型聚合。与文献[21]类似，本地模型训练使用随机梯度下降(Stochastic gradient descent, SGD)算法。当N个用户对均完成E个训练轮次后，本地模型参数发送至边缘服务器进行模型聚合。

边缘服务器为用户$U_1, U_2, \cdots, U_N$各定制一个专用的超网络，超网络为用户输出用于生成个性化语义编码器的聚合权重矩阵$\boldsymbol{\alpha}_i$。超网络的输入$\boldsymbol{\nu}_i$和超网络的参数ψ_i采用链式法则[39]进行更新

$$\Delta\boldsymbol{\nu}_i=(\nabla_{\boldsymbol{\nu}_i}\bar{\theta}_i)\Delta\theta_i=[\{\theta^1,\theta^2,\cdots,\theta^n\}\times\nabla_{\boldsymbol{\nu}_i}HN_i(\boldsymbol{\nu}_i;\psi_i)]^{\mathrm{T}}\Delta\theta_i \tag{8}$$

$$\Delta\psi_i=(\nabla_{\psi_i}\bar{\theta}_i)\Delta\theta_i=[\{\theta^1,\theta^2,\cdots,\theta^n\}\times\nabla_{\psi_i}HN_i(\boldsymbol{\nu}_i;\psi_i)]^{\mathrm{T}}\Delta\theta_i \tag{9}$$

式中：$\Delta\theta_i=\theta_i^{(t+1)}-\theta_i^{(t)}$为用户$U_i$的语义编码器模型参数在本地训练前后的变化，$\nabla$表示求偏导。从式(7,8)可以看出，超网络的输入$\boldsymbol{\nu}_i$和超网络的参数$\psi_i$的更新与用户$U_i$的语义编码器模型参数的更新有关，因此超网络的更新方向与目标网络的更新方向一致。同时，超网络能够捕获目标网络中不同模块之间的贡献程度[39]，因此超网络能够通过输出的矩阵$\boldsymbol{\alpha}_i$有效地反映各个模块重要程度。根据式(7,8)，服务器在每个FL训练轮次更新用户U_i的超网络嵌入向量和参数，然后输出聚合权重矩阵$\boldsymbol{\alpha}_i$。

服务器的超网络聚合模块根据式(6)得到每个用户的个性化语义编码器模型参数。同时，服务器的FedAvg聚合模块采用FedAvg算法聚合用户的信道编码器参数、信道解码器参数以及语义解码器参数，得到全局模型参数。模型参数完成聚合后，服务器将$\bar{\theta}_i$及ϕ_g、χ_g、φ_g发送至各用户。用户根据收到的模型参数更新本地模型参数，启动本地模型训练过程，并进行下一轮的FL训练。具体的DeepSC-FedHN算法流程如算法1所示。

算法1 DeepSC-FedHN算法流程

初始化 总通信轮次T，局部训练轮次E，学习率η，初始化DeepSC-S模型参数θ_i、ϕ_i、χ_i、φ_i，超网络参数ψ_i，超网络嵌入向量$\boldsymbol{\nu}_i$

输入 用户U_i的本地语音数据$\boldsymbol{s}$

(1) while $t<T$ do

用户$U_i(i=1,2,\cdots,N)$:

(2) 初始化本地训练轮次$e=0$

(3) while $e<E$ do

(4) 根据SGD算法更新θ_i、ϕ_i、χ_i、φ_i

(5) end while

(6) 将θ_i、ϕ_i、χ_i、φ_i上传至边缘服务器BS

边缘服务器BS:

(7) for $HN_i(\boldsymbol{\nu}_i;\psi_i)(i=1,2,\cdots,N)$ do

(8) 根据式(5)输出$\boldsymbol{\alpha}_i$

(9) 根据式(6)得到$\bar{\theta}_i$

(11) 根据式(7,8)更新$\boldsymbol{\nu}_i$,ψ_i

(12) end for

(13) FedAvg聚合模块根据FedAvg算法得到ϕ_g、χ_g、φ_g

(14) 边缘服务器将$\bar{\theta}_i$、ϕ_g、χ_g、φ_g发送回各用户

(15) 用户U_i用$\bar{\theta}_i$、ϕ_g、χ_g、φ_g更新模型参数

(16) $t=t+1$

(17) end while

输出 $\bar{\theta}_i$、ϕ_g、χ_g、φ_g

4 实验设置

4.1 联邦实验环境及模型参数设置

设置用户数目为$N=4$,本地模型训练轮次$E=10$,学习率$\eta=0.001$,优化器采用SGD,批处理大小$B=32$,全局通信轮次$T=40$。超网络由3个全连接层组成,超网络的输入嵌入向量$\boldsymbol{\nu}_i$维度为100,学习率为0.000 5,优化器使用SGD。

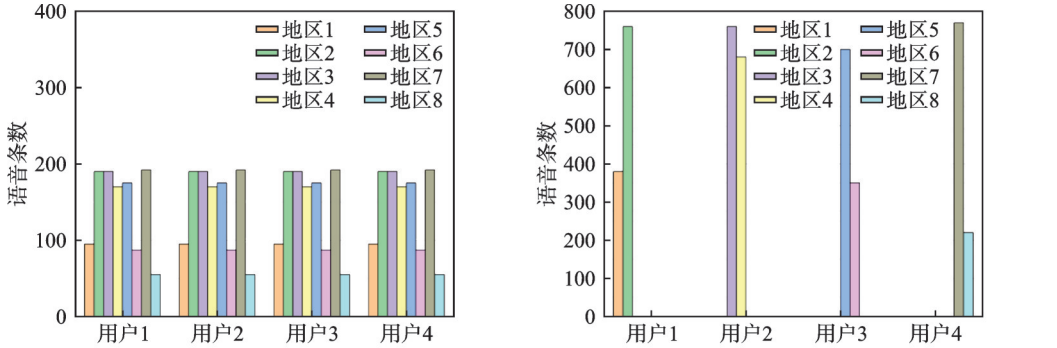
本地模型DeepSC-S模型参数根据文献[21]进行设置,其中语义编码器的SE-ResNet模块数为6,训练和测试过程中的信道类型为高斯信道、瑞利信道和莱斯信道,训练时信道的信噪比(Signal-to-noise, SNR)同文献[21]一致,设置为8 dB,测试时的SNR设置为0~14 dB。实验使用Python中的深度学习工具包Pytorch(版本为1.13.1)进行实现,在1张GeForce RTX 3090 GPU上进行。

4.2 数据集

本文实验数据集采用TIMIT数据集,该数据集的语音采样频率为16 kHz,一共包含6 300条语句,其中训练集包含4 620条语句,测试集包含1 680条语句,由来自美国8个主要方言地区的630个说话人的语句组成,其中每个说话人说出给定的10个句子。对TIMIT数据集分别按照独立同分布(Independent and identically distributed, IID)和Non-IID进行划分。

(1) IID划分方式。用户U_1、U_2、U_3、U_4从训练集的8个地区中分别随机抽取相同数量的语音作为自己的本地训练集,随机地从测试集的8个地区中抽取相同数量的语音作为自己的测试集。各用户抽取的训练数据和测试数据均无重叠。

(2) Non-IID划分方式。用户U_1、U_2、U_3、U_4依次从训练集的8个地区中选择2个地区的语音数据作为自己的本地数据集,对于测试集也进行类似的处理。图4给出了IID划分方式和Non-IID划分方式下的各用户数据分布。



(a) 在IID划分方式下各用户数据分布(TIMIT数据集)
(a) Data distribution of each user under IID partition (TIMIT dataset)

(b) 在non-IID划分方式下各用户数据分布(TIMIT数据集)
(b) Data distribution of each user under non-IID partition (TIMIT dataset)

图4 不同划分方式下各用户数据分布(TIMIT数据集)

Fig.4 Distribution of user data under different partition methods (TIMIT dataset)

同时，本文还使用Edinburgh DataShare语音数据集和THCHS-30语音数据集进行补充实验。Edinburgh DataShare语音数据集的训练集包含了28个说话人的共11 572条英文语音数据，测试集由与训练集说话人不同的2个说话人的共822条英文语音数据组成。对该数据集进行如下划分，将训练集按说话人分为4组，每组包含7个说话人的语音数据，用户$U_i(i=1、2、3、4)$选择其中的1组作为自己的本地数据集进行模型训练。将该数据集的测试集作为每个用户的测试集进行模型测试。

THCHS-30语音数据集是由清华大学语音与语言技术中心出版的开放式中文语音数据库，该数据集的训练集包含了30个说话人的共10 893条中文语音数据，测试集由与训练集说话人不同的10个人的2 496条中文语音数据组成。对该数据集进行如下划分，将训练集按说话人分为4组，每组包含7个说话人的语音数据(其中1组包含9个说话人)，用户$U_i(i=1、2、3、4)$选择其中的一组作为自己的本地数据集进行模型训练。将该数据集的测试集作为每个用户的测试集进行模型测试。

4.3 对比方案

将本文所提出的DeepSC-FedHN方法与本地训练策略、基于FedAvg的训练策略、基于FedProx的训练策略和DeepSC-pFedLA进行了性能对比。对比方案具体描述如下：

(1) 本地训练：所有用户只进行本地训练，不进行全局模型聚合。

(2) FedAvg[26]：采用FedAvg算法对各用户本地训练得到的DeepSC-S模型参数进行聚合更新。

(3) FedProx[37]：采用FedProx算法对各用户本地训练得到的DeepSC-S模型参数进行聚合更新。实验中设置μ值为0.1，μ为控制损失函数中近端项的超参数。

(4) DeepSC-pFedLA[39]：与文献[39]类似，采用超网络为模型的所有层生成聚合权重，根据聚合权重矩阵得到每个用户的个性化DeepSC-S模型。

5 实验结果及分析

5.1 TIMIT数据集实验结果及分析

图5为采用提出的DeepSC-FedHN方法，训练阶段各用户的均方误差(Mean-square error, MSE)损失值与训练轮次的关系。图5中SNR=8 dB，信道为莱斯信道。从图5可以看出，本文所提出的DeepSC-FedHN方法可以为每个用户生成1个稳定的个性化模型，模型大概在400个训练轮次后达到收敛。

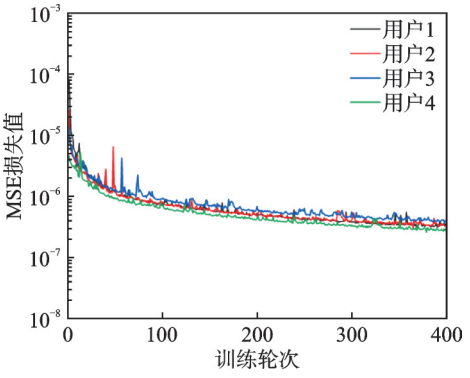


图5 各用户MSE损失值随训练轮次的变化

Fig.5 Variation of MSE loss values of each user with training epochs

图6给出了不同方案在IID划分方式下的TIMIT数据集下的PESQ、SDR和STOI随SNR的变化曲线。从图6可以看出，与本地训练方案、FedAvg方案、FedProx方案相比，在高斯信道、瑞利信道和莱斯信道下，本文提出的DeepSC-FedHN方法的PESQ得分、SDR值以及STOI得分均更高。同时，从图6还可以看出，与DeepSC-pFedLA方案相比，本文提出的DeepSC-FedHN方法在IID划分方式下TIMIT数据集上的性能总体均更优，这说明了在各用户数据具有相同分布的情况下，仅针对提取语义信息的重要模块进行个性化参数更新，能够有效地提升模型的性能。

图7给出了不同方案在Non-IID划分方式下TIMIT数据集的PESQ、SDR和STOI随SNR的变化曲线。从图7中可以看出，与本地训练方案、FedAvg方案、FedProx方案相比，在3种信道下，本文提出

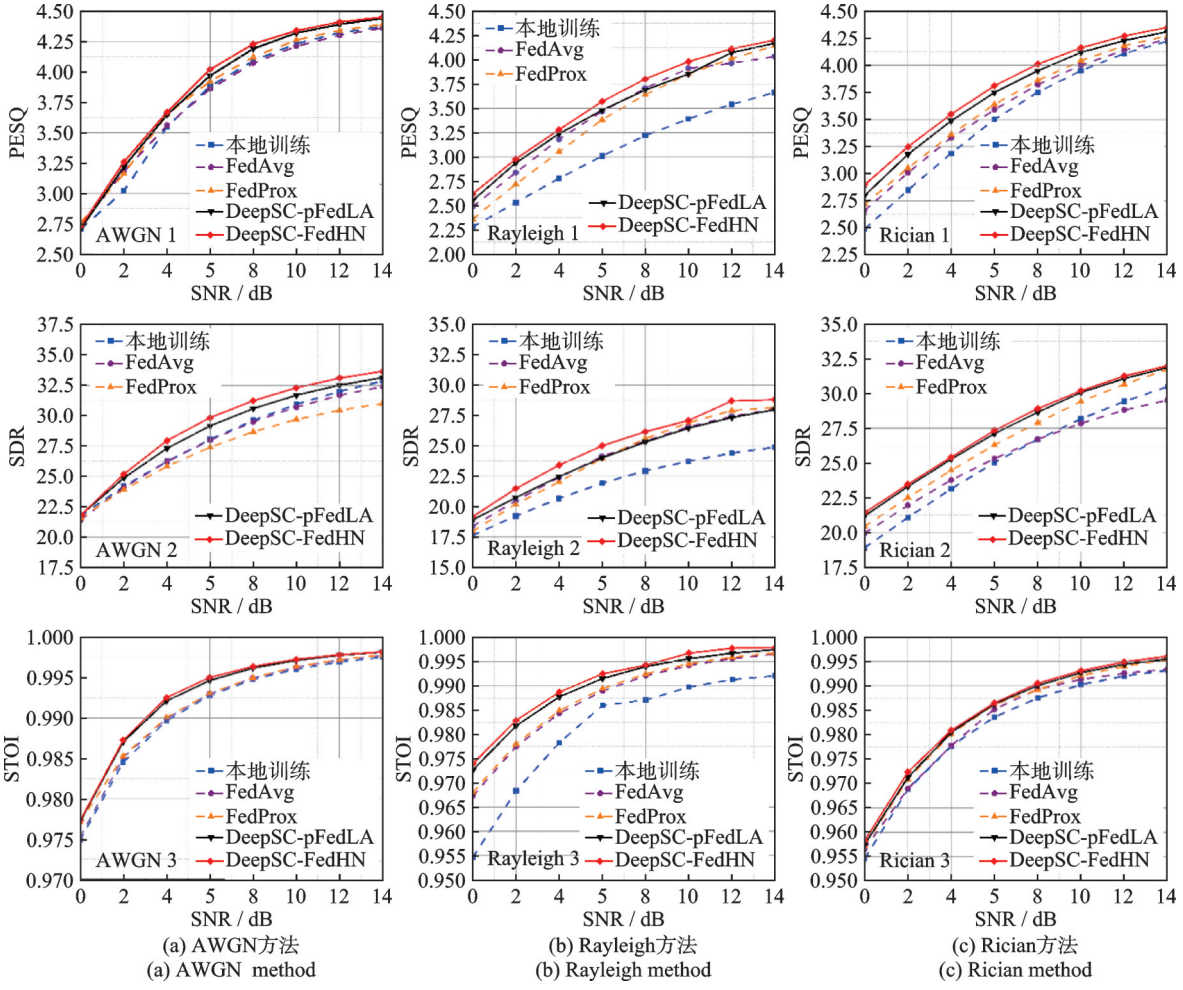


图6 不同方法在IID划分方式下的TIMIT数据集上的PESQ得分、SDR值和STOI得分

Fig.6 PESQ scores, SDR values and STOI scores of different methods on TIMIT datasets divided by IID

的DeepSC-FedHN方法在PESQ、SDR以及STOI上的性能更优。同时,从图7还可以看出,在Non-IID划分方式下的TIMIT数据集上,本文所提出的DeepSC-FedHN方法的PESQ得分、SDR值和STOI得分要低于DeepSC-pFedLA方案。这是因为本文提出的DeepSC-FedHN方法仅对模型的重要模块参数进行更新,而DeepSC-pFedLA对模型所有层参数都进行了更新,得到的模型个性化程度更高,与Non-IID划分方式下数据集的非独立同分布的特点更匹配。但是本文提出的DeepSC-FedHN方法在模型聚合时的计算量远小于DeepSC-pFedLA方法。具体地,根据2.2节中的计算量分析可以得到,在本实验环境下,采用DeepSC-pFedLA方法在模型聚合时的计算量约为1.623×10^{8},采用DeepSC-FedHN模型聚合时的计算量约为4.162×10^{6},仅为前者的2.5%,大幅减少了模型聚合所需的计算量。

5.2 Edinburgh DataShare数据集和THCHS-30数据集实验结果及分析

图8给出了不同方案在在Edinburgh DataShare的语音数据集下的PESQ、SDR和STOI随SNR的变化曲线。与在TIMIT数据集上得到的结果相似,本文提出的DeepSC-FedHN方法与4种对比方案相比,在3种信道下均取得了更高的PESQ得分、SDR值以及STOI得分,这表明本文提出的DeepSC-

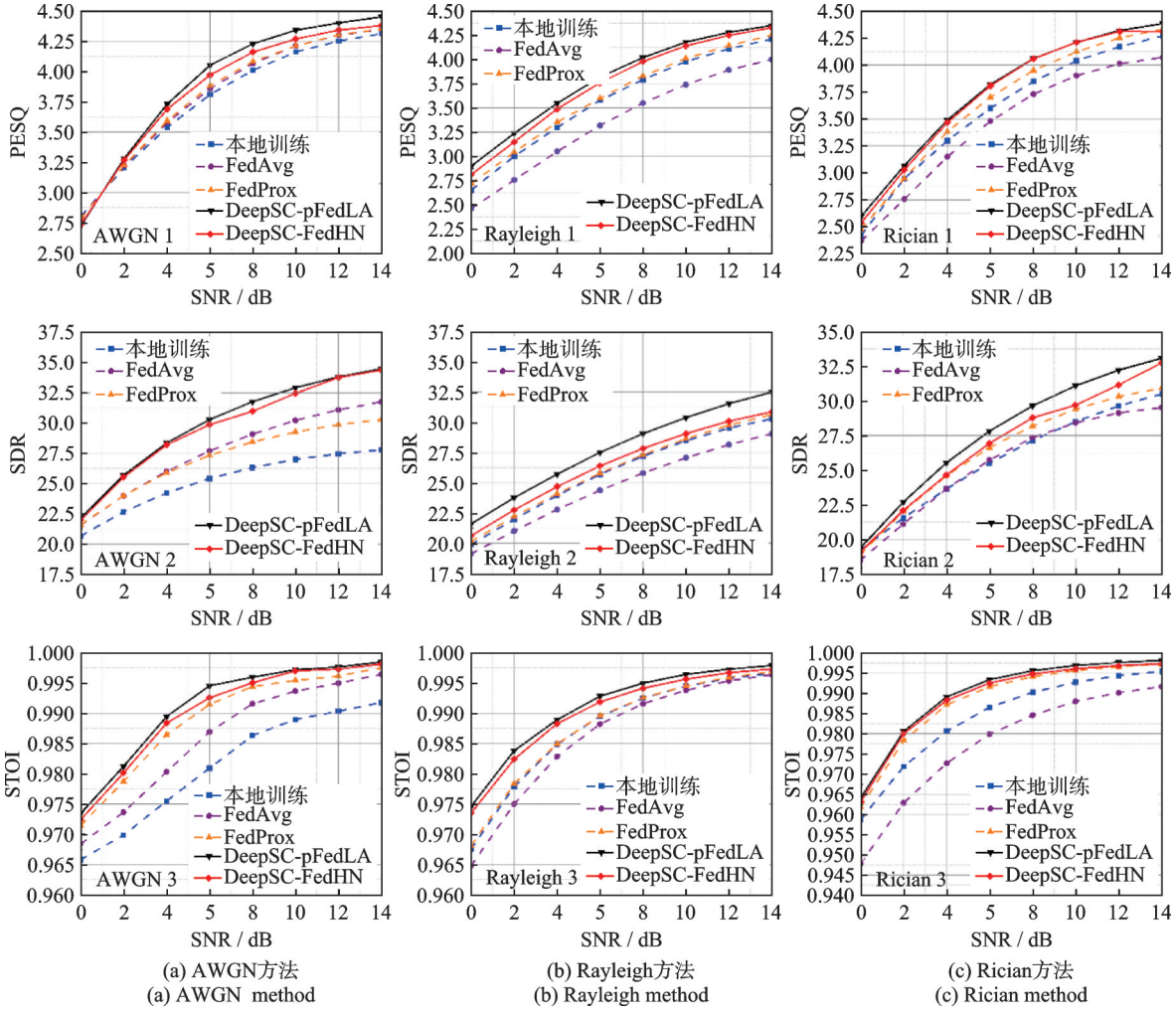


图7 不同方法在Non-IID划分方式下的TIMIT数据集上的PESQ得分、SDR值和STOI得分

Fig.7 PESQ scores, SDR values and STOI scores of different methods on TIMIT datasets divided by Non-IID

FedHN方法具有泛化性。值得提出的是,在Edinburgh DataShare测试集下,本文提出的DeepSC-FedHN性能要优于DeepSC-pFedLA。这是因为该数据集中测试集包含的2个说话人与训练集中的28个说话人都不同,在本实验中相当于模拟了一个新的用户。而本文提出的方法对于语义解码器部分的模型聚合采用FedAvg算法,能够适当降低模型整体的个性化,提高模型对未知数据分布的泛化性,所以当推广到数据分布未知的新用户时,本文提出的方法能够获得更好的性能。

图9给出了不同方案在THCHS-30语音数据集下的PESQ、SDR和STOI随SNR的变化曲线。与在Edinburgh DataShare数据集上得到的结果相似,本文提出的DeepSC-FedHN方法与4种对比方案相比,在3种信道下均取得了更高的PESQ得分、SDR值以及STOI得分,这进一步证明了本文提出的DeepSC-FedHN方法在面对新用户时的泛化性。同时,本文提出的DeepSC-FedHN方法在中文数据集THCHS-30上取得了与在英文数据集Edinburgh DataShare上类似的性能,表明了本文提出的DeepSC-FedHN方法对于不同语种的泛化性。

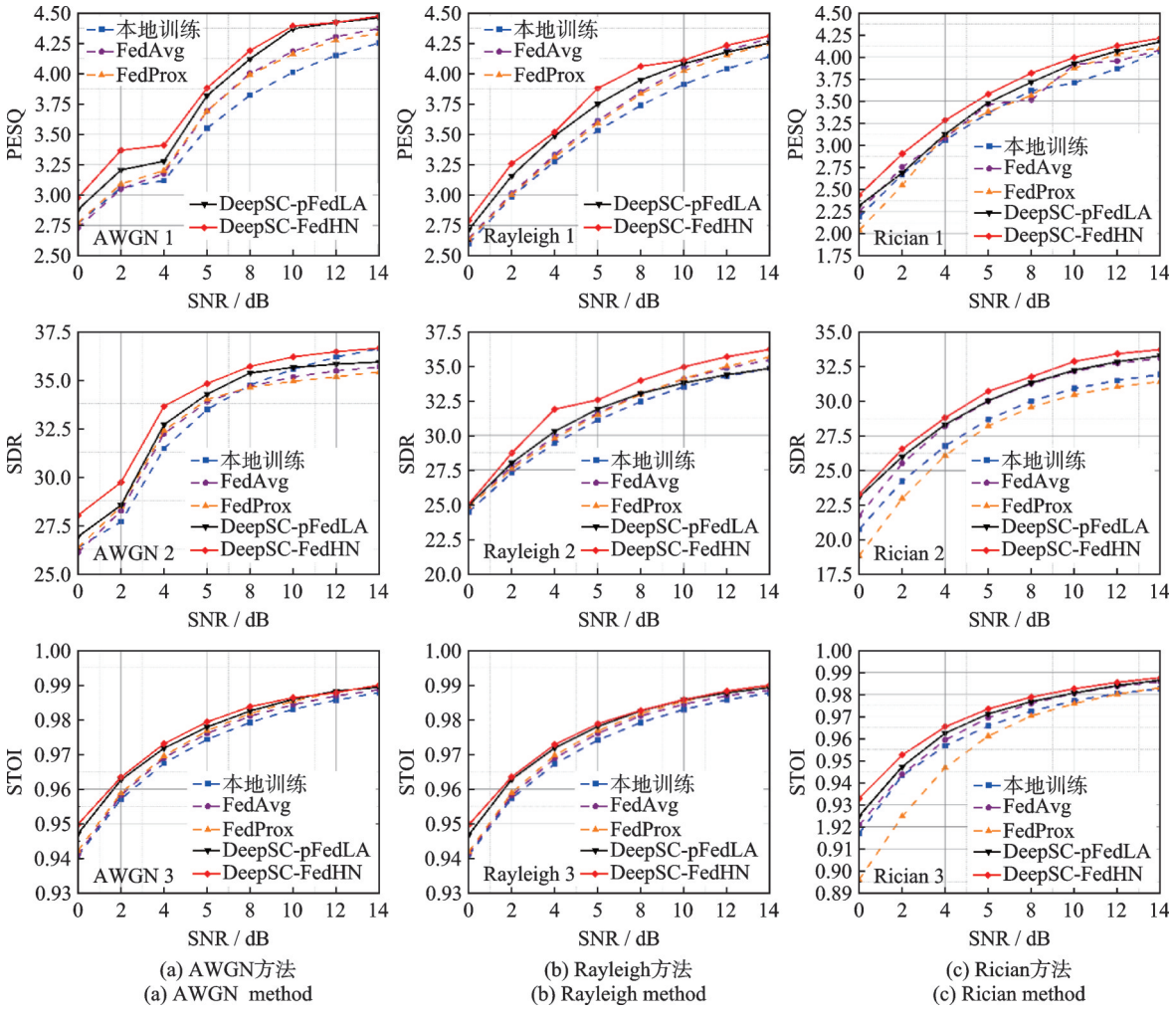


图8 不同方法在Edinburgh DataShare的语音数据集上的PESQ得分、SDR值和STOI得分

Fig.8 PESQ scores, SDR values and STOI scores of different methods on Edinburgh DataShare speech dataset

5.3 模型参数量及推理延迟分析

在模型参数量方面,本文使用的本地模型为DeepSC-S模型,模型参数根据文献[21]进行设置,其中语义编码器和语义解码器的SE-ResNet模块数均为6,模型的参数总量约为2.7×10^6,属于轻量级模型。同时,本文所提DeepSC-FedHN方法在边缘服务器部署的超网络模型由3个全连接层组成,超网络总的参数量约为0.02×10^6,模型参数量远小于本地模型,可近似忽略。因此与本地DeepSC-S模型相比,本文所提的DeepSC-FedHN方法几乎未增加模型的参数量,在本地资源消耗相同的情况下,提升了模型的整体性能。

在模型推理延迟方面,本文使用TIMIT数据集中的测试集对模型的推理时间进行了测试分析。TIMIT数据集中的测试集共有1 680条语句,语音长度为3~7 s,本文所提DeepSC-FedHN方法下本地模型的推理延迟为11.67~12.33 ms,平均推理延迟约为12 ms,远小于语音持续。因此,后续经过合理设计,本地用户能够实现实时高效的语音传输。

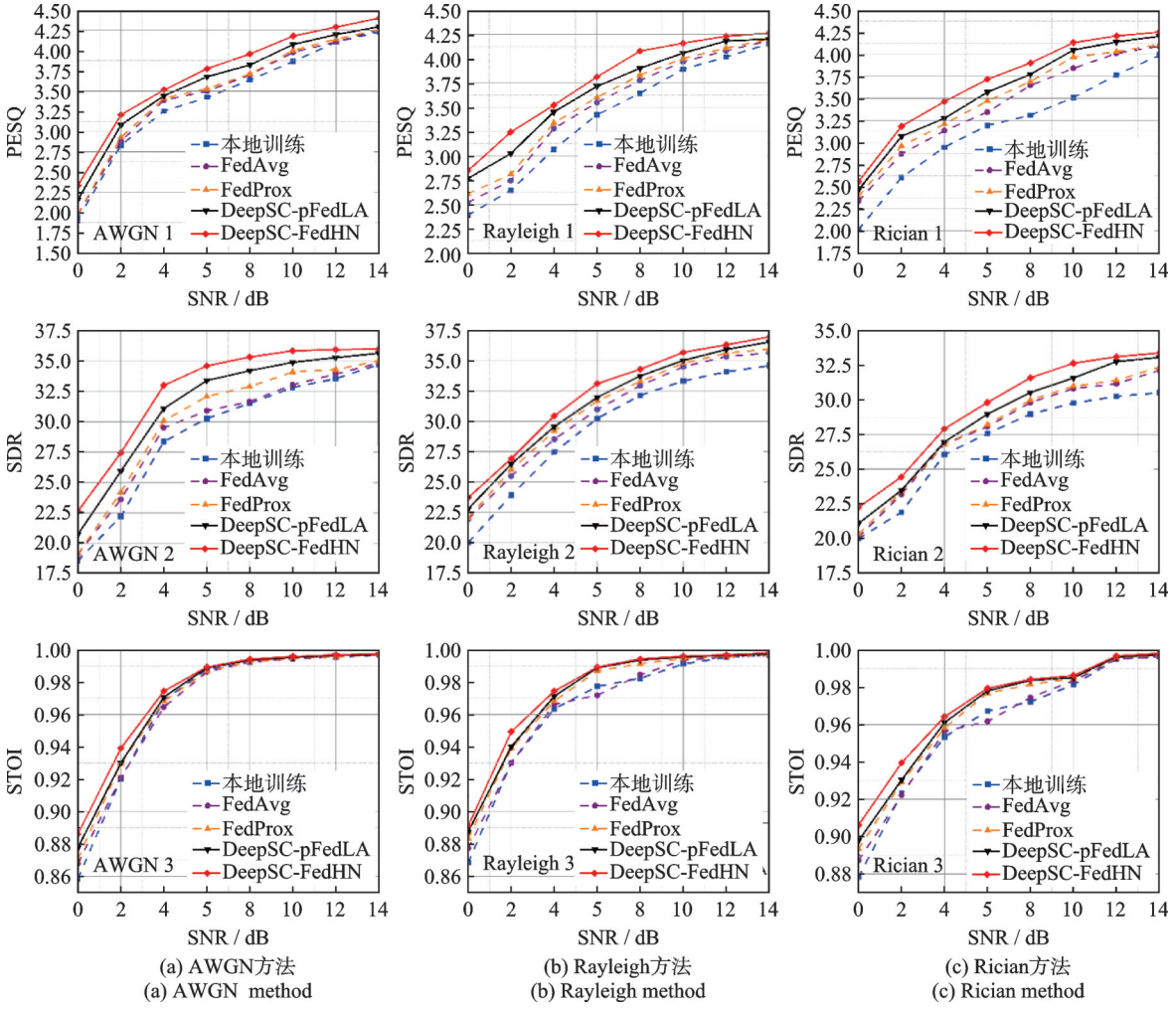


图9 不同方法在THCHS-30语音数据集上的PESQ得分、SDR值和STOI得分

Fig.9 PESQ scores, SDR values and STOI scores of different methods on THCHS-30 speech dataset

6 结束语

本文面向多用户语音语义通信场景提出了DeepSC-FedHN方法，利用超网络为每个用户输出一个个性化的语义编码器，对其他模块使用FedAvg算法，使得模型在具备个性化的同时，提高整体模型的泛化性。在TIMIT数据集、Edinburgh DataShare的语音数据集和THCHS-30数据集上的实验结果表明，与本地训练、FedAvg和FedProx方案相比，所提DeepSC-FedHN方法在PESQ、SDR以及STOI性能指标上有总体上的提升。同时与DeepSC-pFedLA方法相比，本文所提DeepSC-FedHN方法大幅减少了模型聚合计算量，并且在面对未知数据时有更好的泛化性能。

参考文献：

[1] LU Z, LI R, LU K, et al. Semantics-empowered communications: A tutorial-cum-survey[J]. IEEE Communications Surveys & Tutorials, 2024, 26(1): 41-79.

[2] 张平, 牛凯, 姚圣时, 等. 面向未来的语义通信: 基本原理与实现方法[J]. 通信学报, 2023, 44(5): 1-14.
ZHANG Ping, NIU Kai, YAO Shengshi, et al. Semantic communications for future: Basic principle and implementation meth-

odology[J]. Journal on Communications, 2023, 44(5): 1-14.
[3] 石光明, 李莹玉, 谢雪梅. 语义通讯: 智能时代的产物[J]. 模式识别与人工智能, 2018, 31(1): 91-99.
SHI Guangming, LI Yingyu, XIE Xuemei. Semantic communications: Outcome of the intelligence era[J]. Pattern Recognition and Artificial Intelligence, 2018, 31(1): 91-99.
[4] MAO J, XIONG K, LIU M, et al. A GAN-based semantic communication for text without CSI[EB/OL]. (2023-12-28) [2024-09-28]. https://arxiv.org/abs/2312.16909.
[5] GETU T, KADDOUM G, BENNIS M. Deep learning-enabled text semantic communication under interference: An empirical study[EB/OL]. (2023-10-30) [2024-09-28]. https://arxiv.org/abs/2310.19974.
[6] 佟国香, 董田荣, 胡珩彰. 视觉注意与语义感知联合推理实现场景文本识别[J]. 数据采集与处理, 2023, 38(3): 665-675.
TONG Guoxiang, DONG Tianrong, HU Hengzhang. Joint inference of visual attention and semantic perception for scene text recognition[J]. Journal of Data Acquisition and Processing, 2023, 38(3): 665-675.
[7] TANG S, LIU C, YANG Q, et al. Secure semantic communication for image transmission in the presence of eavesdroppers [EB/OL]. (2024-04-18) [2024-09-28]. https://arxiv.org/abs/2404.12170.
[8] 陈善学, 王程. 融合多时间维度视觉与语义信息的图像描述方法[J]. 数据采集与处理, 2024, 39(4): 922-932.
CHEN Shanxue, WANG Cheng. Image captioning method for fusing multi-temporal dimensional visual and semantic information[J]. Journal of Data Acquisition and Processing, 2024, 39(4): 922-932.
[9] 何晨光, 黄声显, 陈舒怡, 等. 基于语义通信的低比特率图像语义编码方法[J]. 信号处理, 2023, 39(3): 410-418.
HE Chenguang, HUANG Shengxian, CHEN Shuyi, et al. A low bitrates image semantic coding method based on semantic communication[J]. Journal of Signal Processing, 2023, 39(3): 410-418.
[10] GAO H, SUN M, XU X, et al. Semantic communication-enabled wireless adaptive panoramic video transmission[EB/OL]. (2024-06-23) [2024-09-28]. https://arxiv.org/abs/2402.16581.
[11] ZHANG Z, YANG Q, HE S, et al. Deep learning enabled semantic communication systems for video transmission[C]//Proceedings of 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall). [S.l.]: IEEE, 2023: 1-5.
[12] JIANG F, PENG Y, DONG L, et al. Large AI model empowered multimodal semantic communications[EB/OL]. (2024-08-04) [2024-09-28]. https://arxiv.org/abs/2309.01249.
[13] ZHANG G, HU Q, QIN Z, et al. A unified multi-task semantic communication system for multimodal data[J]. IEEE Transactions on Communications, 2024, 72(7): 4101-4116.
[14] SHI G, XIAO Y, LI Y, et al. From semantic communication to semantic-aware networking: Model, architecture, and open problems[J]. IEEE Communications Magazine, 2021, 59(8): 44-50.
[15] WENG Z, QIN Z, LI G. Semantic communications for speech recognition[C]//Proceedings of 2021 IEEE Global Communications Conference (GLOBECOM). [S.l.]: IEEE, 2021: 1-6.
[16] WENG Z, QIN Z, TAO X. Semantic-aware speech-to-text transmission over MIMO channels[C]//Proceedings of 2023 IEEE International Conference on Communications Workshops (ICC Workshops). [S.l.]: IEEE, 2023: 1362-1367.
[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.
[18] HAN T, YANG Q, SHI Z, et al. Semantic-preserved communication system for highly efficient speech transmission[J]. IEEE Journal on Selected Areas in Communications, 2022, 41(1): 245-259.
[19] WENG Z, QIN Z. Robust semantic communications for speech transmission[EB/OL]. (2024-04-25) [2024-09-28]. https://arxiv.org/abs/2403.05187.
[20] WENG Z, QIN Z, TAO X. Task-oriented semantic communications for speech transmission[C]//Proceedings of 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall). [S.l.]: IEEE, 2023: 1-5.
[21] WENG Z, QIN Z. Semantic communication systems for speech transmission[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2434-2444.
[22] XIAO Z, YAO S, DAI J, et al. Wireless deep speech semantic transmission[C]//Proceedings of ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.
[23] WENG Z, QIN Z, TAO X, et al. Deep learning enabled semantic communications with speech recognition and synthesis[J]. IEEE Transactions on Wireless Communications, 2023, 22(9): 6227-6240.
[24] WEI H, XU W, WANG F, et al. SemAudio: Semantic-aware streaming communications for real-time audio transmission[C]//

Proceedings of GLOBECOM 2022—2022 IEEE Global Communications Conference. [S.l.]: IEEE, 2022: 3965-3970.

[25] GRASSUCCI E, MARINONI C, RODRIGUEZ A, et al. Diffusion models for audio semantic communication[C]// Proceedings of ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2024: 13136-13140.

[26] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]//Proceedings of Artificial Intelligence and Statistics. [S.l.]: PMLR, 2017: 1273-1282.

[27] TONG H, YANG Z, WANG S, et al. Federated learning for audio semantic communication[J]. Frontiers in Communications and Networks, 2021, 2: 734402.

[28] DENG Z, LI S, CAI Y, et al. Federated learning for image semantic communication system based on CNN and Transformer [C]//Proceedings of 2023 International Conference on Ubiquitous Communication (UCom). [S.l.]: IEEE, 2023: 408-414.

[29] XIE B, WU Y, SHI Y, et al. Communication-efficient framework for distributed image semantic wireless transmission[J]. IEEE Internet of Things Journal, 2023, 10(24): 22555-22568.

[30] WEI H, NI W, XU W, et al. Federated semantic learning driven by information bottleneck for task-oriented communications [J]. IEEE Communications Letters, 2023, 27(10): 2652-2656.

[31] XIE R, LI C, ZHOU X, et al. Asynchronous federated learning for real-time multiple licence plate recognition through semantic communication[C]//Proceedings of ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2023: 1-5.

[32] 涂勇峰, 陈文. 基于联邦学习的多用户语义通信系统部署方法[J]. 信号处理, 2022, 38(12): 2486-2495.
TU Yongfeng, CHEN Wen. A deployment method of multi-user semantic communication system based on federated learning [J]. Journal of Signal Processing, 2022, 38(12): 2486-2495.

[33] ZHANG J, QU Z, CHEN C, et al. Edge learning: The enabling technology for distributed big data analytics in the edge[J]. ACM Computing Surveys (CSUR), 2021, 54(7): 1-36.

[34] FALLAH A, MOKHTARI A, OZDAGLAR A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach[J]. Advances in Neural Information Processing Systems, 2020, 33: 3557-3568.

[35] HUANG Y, CHU L, ZHOU Z, et al. Personalized cross-silo federated learning on non-IID data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2021: 7865-7873.

[36] HA D, DAI A, LE Q. Hypernetworks[EB/OL]. (2016-12-01) [2024-09-28]. https://arxiv.org/abs/1609.09106.

[37] LI T, SAHU A, ZAHEER M, et al. Federated optimization in heterogeneous networks[J]. Proceedings of Machine learning and systems, 2020, 2: 429-450.

[38] SHAMSIAN A, NAVON A, FETAVA E, et al. Personalized federated learning using hypernetworks[C]//Proceedings of International Conference on Machine Learning. [S.l.]: PMLR, 2021: 9489-9502.

[39] MA X, ZHANG J, GUO S, et al. Layer-wised model aggregation for personalized federated learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2022: 10092-10101.

[40] MOLCHANOV P, TYREE S, KARRAS T, et al. Pruning convolutional neural networks for resource efficient inference[EB/OL]. (2017-06-08) [2024-09-28]. https://arxiv.org/abs/1611.06440.

作者简介:

刘月照(2000-),男,硕士研究生,研究方向:语义通信、智能语音信号处理,E-mail: 1222014729@njupt.edu.cn。

郭海燕(1983-),**通信作者**,女,博士,副教授,研究方向:语义通信、智能语音信号处理和RIS辅助无线通信,E-mail:guohy@njupt.edu.cn。

王添顺(1998-),男,博士,讲师,研究方向:联邦学习、边缘智能以及安全计算,E-mail:tswang@njupt.edu.cn。

陈飞飞(1999-),男,硕士研究生,研究方向:智能语音信号处理,E-mail:1223014339@njupt.edu.cn。

(编辑:刘彦东)

Speech Transmission System Based on Personalized Federated Learning and Semantic Communication

LIU Yuezhao, GUO Haiyan*, WANG Tianshun, CHEN Feifei

(School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: In multi-user speech transmission scenarios, the statistical heterogeneity of data among different users results in the transmission performance degradation if a uniform semantic communication based speech transmission model is used by all users. To address this problem, this paper proposes a novel deep learning-based semantic communication system using federated learning based on hypernetworks (DeepSC-FedHN), enabling each user to obtain a personalized model adaptive to its own data characteristics without compromising data privacy. Specifically, considering that different modules of the semantic encoder play different roles in extracting semantic information, the edge server employs a per-user hypernetwork to generate a personalized aggregation weight matrix by dynamically evaluating the importance of each module in the semantic encoder. The generated aggregation weight matrix is then used to update the corresponding model parameters, effectively tailoring the global knowledge to different users' needs. Concurrently, since the channel codec and semantic decoder are not involved in extracting the semantic features of each local users' data, the standard federated averaging (FedAvg) algorithm is used to perform weighted aggregation and updates on the channel codecs and semantic decoders of all the users. Experimental results on TIMIT and Edinburgh DataShare datasets show that the proposed DeepSC-FedHN scheme leads to significant improvement of speech transmission performance. Specifically, it outperforms conventional local training, the standard FedAvg approach, the federated proximal (FedProx) method, and the layer-wise personalized FL scheme (DeepSC-pFedLA) in terms of perceptual evaluation of speech quality (PESQ), signal-to-distortion ratio (SDR) and short time objective intelligibility (STOI), particularly in non-independent and identically distributed (non-IID) data settings. Additionally, the proposed DeepSC FedHN model exhibits better generalization ability for unseen speakers' data and also demonstrates significantly lower computational overhead for model aggregation compared to the DeepSC pFedLA. We conclude that the integration of a hypernetwork for generating personalized weights offers a highly effective mechanism for tackling data heterogeneity in federated semantic communication systems, leading to superior and more adaptable speech transmission performance while fully preserving user data privacy.

Highlights:

1. Propose a novel deep learning-based semantic communication system using federated learning based on hypernetworks (DeepSC-FedHN) for personalized multi-user speech semantic communication.

2. Use a hypernetwork to generate user-specific aggregation weights for encoders and aggregates channel codec and decoder via standard federated learning.

3. Achieve superior performance under non-IID data while preserving privacy.

Key words: speech transmission; semantic communication; federated learning; hypernetwork

---

Foundation item: National Natural Science Foundation of China (No.62071242)。
Received: 2024-09-29; **Revised**: 2025-02-18
***Corresponding author, E-mail:** guohy@njupt.edu.cn.