北京交通大学

硕士学位论文

基于时空图卷积网络的多通道语音增强算法研究

Research on Multichannel Speech Enhancement Algorithm based on Spatial-Temporal Graph Convolutional Network

作者：郝明辉

导师：余晶晶

北京交通大学

2022 年 6 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：郝明辉　　导师签名：

签字日期：2022年6月2日　　签字日期：2022年6月2日

学校代码：10004 密级：公开

北京交通大学

硕士学位论文

基于时空图卷积网络的多通道语音增强算法研究

Research on Multichannel Speech Enhancement Algorithm based on Spatial-Temporal Graph Convolutional Network

作者姓名：郝明辉 学　　号：19120008

导师姓名：余晶晶 职　　称：副教授

学位类别：工学 学位级别：硕士

学科专业：电子科学与技术 研究方向：多维信号处理

北京交通大学

2022 年 6 月

致谢

本文的工作是在余晶晶老师的悉心指导下完成的，真诚感谢余老师在整个研究生阶段对我的指导和帮助。余老师学术态度严谨认真，科研探索一丝不苟，引导我建立了系统的思维方式、完善的逻辑结构、科学的研究方法。在余老师的培养下我获得了非常多的成长，衷心感谢余老师。

感谢实验室的师兄师姐师弟师妹，李润雷、奚琦、刘兴春、王鸿旭、何淇、陈广磊、张璐瑶和孙志明，这段经历有你们的陪伴而更加丰富多彩。特别感谢奚琦师姐、何淇师妹和张璐瑶师妹在完成这篇论文过程中提供的大力支持。

人生的许多阶段都充满了随机性，能走到今天我深深地感谢初三班主任张留柱老师，不仅感谢他在学业上对我无怨无悔的无私付出，更感谢他在我人生的迷茫期为我指明了一条正确的路。

感谢我的父母，两位人生导师帮助我建立起朴素善良的价值观，一直以来给我极大的自由，永远在背后默默地付出，从不对我有过多要求，辛苦操劳半生只希望我能活得不那么累。感谢父母给予我的这份世间最清澈的爱。

摘要

多通道语音增强算法利用麦克风阵列捕捉信号的时域、空域、频域等多重特征来进行目标和非目标信号分量的分离与估计，从而抑制干扰和噪声，提升陌生多变场景下的目标语音增强效果。近年来，相比于传统的波束成形方法，基于机器学习的数据驱动算法在语音增强领域得到了有效应用，尤其是在应对突发噪声方面取得了显著地性能提升。然而，当前的多通道语音增强算法仍然没有解决以下几个具有较大挑战的瓶颈问题：(1) 仅简单挪用机器视觉和自然语言处理领域的现有网络，没有结合声阵列信号的独特机理来进行网络模型的高效性、可控性和可解释性改造，尚无法适用于小型智能人机交互设备；(2) 由于缺乏准确的阵列分布先验信息，难以对多通道信号间的空间关联关系进行完全解析，导致实际应用中的声源盲分离效果不佳；(3) 没有针对声阵列信号的复杂时空关联关系进行有效建模和解相关，从而加剧目标与非目标信号分量的混叠。因此，本文拟针对声阵列信号中所隐含的复杂时空关联关系进行非欧空间的图理论建模，构建基于通道相关性和语音信号时频关联性的图聚合运算和动态邻接矩阵；搭建时空图卷积语音增强网络，在缺失阵列和场景准确先验信息的情况下，显著提升目标语音增强质量；提出复频谱网络拓展方法、模块参数量优化方法和基于语音可懂度的目标重建损失函数三种网络优化策略，进一步为算法在小型智能人机交互设备的落地应用提供可行方案和技术支持。具体研究内容如下：

(1) 在缺失阵列和场景先验信息的情况下，基于非欧空间图理论，进行多通道语音增强问题的图建模，解析声阵列信号中所隐含的与阵列拓扑和声源位置相关的空间关联关系。具体包括：针对麦克风时空数据的图结构建模、多通道信号的时空频图聚合运算设计、可反映长时通道关联关系的动态自适应邻接矩阵构建和图神经网络搭建。

(2) 融合基于图卷积运算的空间关联性提取方法和基于时频卷积的时域关联性提取方法，构建时空图卷积语音增强网络，提取声阵列信号的时、空、频关联特征进行多源信号分离，在缺失场景和阵列先验信息的情况下，显著提升目标语音重建质量，抑制噪声和干扰。实验证明，本文提出的时空图卷积语音增强网络相比目前最优的算法在多种噪声场景下取得了超过 11%的语音质量感知评估性能的提升，且主观评价指标也达到了最优效果。

(3) 为了进一步解决语音增强网络应用于小型智能人机交互设备所面临的瓶颈，针对目标语音重建的相位信息损失问题，进行了基于复频谱的时空图卷积语音增强网络拓展；进行了网络参数量和系统实时性的优化；进行了基于人耳听觉感知的语音可懂度网络训练损失函数设计。实验证明，基于复频谱的时空图卷积语音

增强网络提升了目标语音增强性能的上限，网络参数量和实时性优化为算法的工程落地提供了技术支持，基于语音可懂度的损失函数可使网络输出的重建目标语音更符合人耳听觉感知。

关键词：空间关联性；声阵列信号图聚合；时空关联性解析与融合；时空图卷积网络；多通道语音增强

ABSTRACT

The multi-channel speech enhancement algorithm uses the microphone array to capture multiple features of the signal in the time domain, spatial domain, and frequency domain to separate and estimate the target and non-target signal components, suppressing interference and noise and improving the target speech enhancement effect in unfamiliar and variable scenes. In recent years, data-driven algorithms based on machine learning have been effectively applied in the field of speech enhancement compared with traditional beamforming methods, and have achieved significant performance improvements, especially in dealing with burst noise. However, current multichannel speech enhancement algorithms still do not address the following bottlenecks, which pose a greater challenge: (1) It simply appropriating existing machine learning networks without combining the unique mechanism of acoustic array signals for efficient, controllable, and interpretable modification of the network model, which is not yet applicable to small intelligent human-computer interaction devices. (2) Due to the lack of accurate array distribution a priori information, it is difficult to fully exploited the spatial correlation between multi-channel signals, resulting in poor blind separation of sound sources in practical applications. (3) No effective modeling and decorrelation of the complex spatial-temporal correlations of the acoustic array signals, thus exacerbating the confounding of target and non-target signal components. Therefore, this paper intends to model the graph theory in non-Euclidean space for the complex spatial-temporal correlations implied in the acoustic array signals, and construct a graph aggregation operation and dynamic adjacency matrix based on channel correlation and time-frequency correlation of speech signals. Build a spatial-temporal graph convolutional speech enhancement network to significantly improve the quality of target speech enhancement in the absence of accurate a priori information about arrays and scenes. Three network optimization strategies, namely, the complex spectrum network expansion, the module parameter optimization method and the target reconstruction loss function based on speech intelligibility index, are proposed to further provide feasible solutions and technical support for the implementation of the algorithm in small intelligent human-computer interaction devices. The details of the study are as follows:

(1) In the absence of array and scene a priori information, graph modeling of the multichannel speech enhancement problem is performed based on non-Euclidean space

graph theory to parse the spatial correlations implied in the acoustic array signal related to the array topology and source location. Specifically, it includes: graph structure modeling for microphone spatial-temporal data, multi-channel spatial-temporal graph aggregation operation design, dynamic adaptive adjacency matrix construction that can reflect long-time channel association relations and graph neural network construction.

(2) The spatial correlation extraction method based on graph convolution operation and the time-domain correlation extraction method based on time-frequency convolution are fused to construct a spatial-temporal graph convolution speech enhancement network to extract the temporal, spatial and frequency correlation features of acoustic array signals for multi-source signal separation, which significantly improves the quality of target speech reconstruction and suppresses noise and interference in the absence of scene and array a priori information. Experiments demonstrate that the spatial-temporal graph convolution speech enhancement network proposed in this paper achieves more than 11% performance improvement in speech quality perception evaluation compared with the current optimal algorithm in a variety of noisy scenarios, and the subjective evaluation metrics also achieve optimal results.

(3) In order to further solve the bottleneck of speech enhancement network applied to small intelligent human-computer interaction devices, a complex spatial-temporal graph convolution speech enhancement network extension is carried out to address the phase information loss problem of target speech reconstruction. Optimization of the number of network parameters and real-time system performance. Designed a loss function based on speech intelligibility of human ear auditory perception. Experiments demonstrate that the spatial-temporal graph convolution speech enhancement network based on the complex spectrum improves the upper limit of the target speech enhancement performance, the network parameters number and real-time optimization provide technical support for the engineering implementation of the algorithm, and the loss function based on the speech intelligibility can make the reconstructed target speech output from the network more consistent with the human ear auditory perception.

KEYWORDS：Spatial dependency; Acoustic array signal graph aggregation; Spatial and temporal correlation analysis and fusion; Spatial-temporal graph convolution network; Multichannel speech enhancement

目录

1 引言

1.1 研究背景及意义

语音是人们生活中获取和传递信息的重要载体[1]。语音信号处理是信号处理领域的一个重要分支，其研究内容包括了语音增强、语音识别、语音合成、目标源分离、语音编码等。近年来随着数字信号处理理论的发展完善和硬件资源运算速度的飞速提升，语音信号处理技术在各研究方向逐步深入，已经有许多成熟的智能语音交互产品融入到我们的日常生活中，如声纹锁、智能音箱、录音笔等。这些产品往往需要先对语音信号进行增强处理，去除输入信号中含有的噪声信息，以提高语音交互质量和识别率。另外，在一些专业领域如养老医疗，通过导诊机器人等人机交互手段实现医生与患者的有效沟通，或使用混合现实设备实现多人多地实时沟通等，都需要对场景中的含噪语音信号进行增强，抑制背景噪声和干扰，以获得更好的交互体验。

图 1-1 智能语音交互设备

Figure 1-1 Intelligent voice interaction devices

语音增强是对麦克风接收到的混合音频信号中的背景噪声和干扰进行抑制，恢复出相对干净的目标说话人语音的技术，其利用干净语音信号和噪声信号具有的不同特征进行目标与非目标信号分量的分离，来获取干净语音。例如干净语音信号在时域具有稳定的包络结构、在频域有明显的频谱分布，而噪声信号中的平稳噪声在时域不明显表现出包络结构，非平稳噪声在频域具有连续的频率分布。语音增强技术的目标是提升重建目标语音的信号干扰比(Signal-to-Interference Ratio, SIR)、信号噪声比(Signal-to-Noise Ratio, SNR)和语音可懂度。

语音增强技术按照麦克风接收信号的通道数可分为单通道语音增强和多通道语音增强。单通道语音增强算法利用信号的时域和频域信息实现目标语音增强。

经典的单通道语音增强算法有谱减法[2]、维纳滤波法[3]、基于统计模型的方法[4]等。近年来基于深度神经网络(Deep Neural Network, DNN)的方法逐渐成为研究主流，包括卷积神经网络(Convolutional Neural Network, CNN)[5]、循环神经网络(Recurrent Neural Network, RNN)[6]、卷积循环神经网络(Convolutional Recurrent Network, CRN)[7]、注意力机制(Attention Mechanism)[8]和生成对抗网络(Generative Adversarial Network, GAN)[9]等方法。单通道语音增强虽然只需单个麦克风即可完成信号的采集，对硬件的成本要求较低，但无法利用声源的空间信息，并且在真实场景中，信号传播由于受墙面等物体发生反射而具有多径效应，造成含噪信号存在着时间和频谱上的混叠，使得单通道语音增强算法性能受限。而多通道语音增强算法不仅能提取信号的时域和频域信息，还能根据信号到达阵列中不同阵元的时间差和声信号传播衰减函数进行声源信号的空间特征提取。例如人耳可看作一个简易的多通道系统，在接收空间音频过程中利用不同声源到达双耳的时延差和声级差进行空间位置确定，利用掩蔽效应实现对干扰噪声信号的过滤。所以多通道语音增强算法可以在真实场景中对干扰和背景噪声达到更有效的抑制，取得较好的 SIR 和 SNR。市场上主流的高品质智能音箱等人机交互产品，往往也是内置多个麦克风组成的麦克风阵列进行多通道语音增强。因此，本文研究的重点放在多通道语音增强算法中。

多通道语音增强可分为基于波束成形的语音增强算法、基于盲源分离的语音增强算法和基于 DNN 的语音增强算法。基于波束成形的语音增强算法根据麦克风阵列拓扑及阵列接收到的实时数据对每个通道信号进行加权估计，获得增强的目标语音信号。但由于其受到阵列结构约束、信号传播模型失配等先验信息不足或者存在误差的影响，在实际应用中，波束成形算法往往无法达到令人满意的效果，在某些场景下甚至会出现严重性能恶化。基于盲源分离的多通道语音增强算法往往会对信号及其分布进行独立性假设，现实场景中信号往往不能完全满足这些假设，比如麦克风在接收不同频率的噪声信号与语音信号时会出现频谱混叠，因此会影响其语音增强性能上限，且无法适用于动态多源场景。基于 DNN 的语音增强算法通过对大量数据进行训练并以干净语音为标签实现有监督的语音增强，近年来取得了较为显著的性能提升，尤其在应对突发干扰方面。但是此类方法从纯数据驱动的角度提取干净语音特征和噪声特征，而忽视了针对阵列采集信号的空间关联性、时-空依赖性的分析和利用，且大多简单挪用机器视觉和自然语言处理领域的现有网络，不适用于要求小型化和实时性的便携智能语音交互设备。

综上所述，本文拟针对阵列结构未知的陌生多变场景，构建可解析多通道混合信号复杂时空关联关系和阵列拓扑特征的非欧空间图理论模型和动态邻接矩阵，提出时-空-频信息融合的时空图卷积语音增强网络，提取多通道含噪信号中的目

标成分，抑制噪声和干扰成分，实现无需先验信息的鲁棒的多通道语音增强。本文的研究成果可广泛应用于视频会议、智慧家庭、车载人机交互系统、三维沉浸式感知系统等场景中。

1.2 多通道语音增强研究现状

1.2.1 基于波束成形的语音增强算法

基于波束成形的语音增强算法的核心思想是根据麦克风阵列及一定的准则设计波束成形器对声阵列信号进行空间滤波，即将多通道信号按照一定的权重进行叠加，从而重建目标信号分量，如图 1-2 所示。

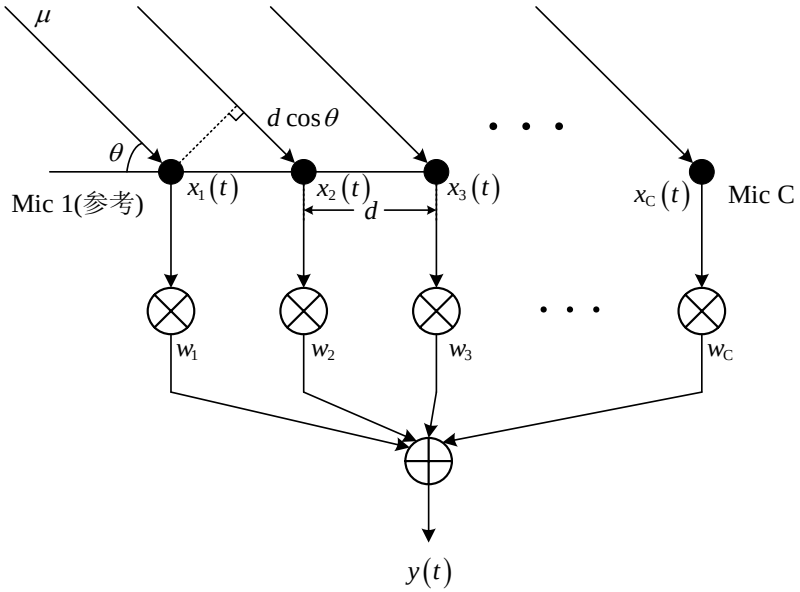


图 1-2 波束成形器的原理示意图

Figure 1-2 Schematic diagram of beamformer principle

假设阵列接收到的多通道含噪语音信号为 $\boldsymbol{x}(t)$，则波束成形器的输出可表示为：

$$y(t)=\boldsymbol{w}^{\mathrm{H}}\boldsymbol{x}(t) \tag{1-1}$$

其频域表达为：

$$\boldsymbol{Y}(f)=\boldsymbol{\mathcal{W}}^{\mathrm{H}}(f)\boldsymbol{X}(f) \tag{1-2}$$

其中，$\boldsymbol{\mathcal{W}}(f)=\left[\mathcal{W}_1(f),\mathcal{W}_2(f),...,\mathcal{W}_{\mathrm{C}}(f)\right]^{\mathrm{T}}$ 为波束成形器的通道加权向量，$\boldsymbol{X}(f)=\left[\boldsymbol{X}_1(f),\boldsymbol{X}_2(f),...,\boldsymbol{X}_{\mathrm{C}}(f)\right]^{\mathrm{T}}$ 为阵列接收到的含噪多通道信号的频域表达，C 为麦克风个数，H 表示共轭转置，T 表示转置。

根据$\boldsymbol{\mathcal{W}}$的求解方法可以将波束成形语音增强算法分为固定波束成形算法和自适应波束成形算法。固定波束成形算法中$\boldsymbol{\mathcal{W}}$与麦克风阵列结构有关，而自适应波束成形算法中，$\boldsymbol{\mathcal{W}}$除了与麦克风阵列结构有关外，还与当前时刻阵列信号本身有关。其中，延迟求和波束成形算法(Delay and Sum Beamforming, DSB)[10]是最常见的固定波束成形算法，其加权向量由声源信号到达麦克风阵列的时间差确定，以阵列中一个麦克风为参考，信号到达第c个麦克风相比于到达参考麦克风的时间差为：

$$\tau_c = \frac{d_c \cos\theta}{\mu} \tag{1-3}$$

其中，d_c为第c个麦克风到第 1 个麦克风的距离，μ为声波传播速度。因此基于DSB 波束成形器的通道加权向量表示如下：

$$\boldsymbol{\mathcal{W}}(f) = \left[1, e^{-j2\pi f\tau_1}, e^{-j2\pi f\tau_2}, ..., e^{-j2\pi f\tau_{C-1}}\right]^{\mathrm{T}} \tag{1-4}$$

自适应波束成形算法(Adaptive Beamforming)根据不同的准测进行$\boldsymbol{\mathcal{W}}$的求解。其中，最小方差无失真响应(Minimum Variance Distortion-less Response, MVDR)波束成形算法[11]是理论最优、也是当前应用最广泛的自适应波束成形算法。该算法基于最小均方误差(Minimum Mean Square Error, MMSE)准则，核心思想是使波束成形器在声源目标方向增益最大，总输出能量最小，从而完成对干扰及噪声分量的抑制，提升 SNR 和 SIR。MVDR 方法的设计准则可以表示为：

$$\begin{aligned} &\min_{w} \quad \boldsymbol{\mathcal{W}}^{\mathrm{H}}(f)\boldsymbol{\mathcal{R}}^{(i+n)}(f)\boldsymbol{\mathcal{W}}(f), \\ &s.t. \qquad \boldsymbol{\mathcal{W}}^{\mathrm{H}}(f)\boldsymbol{r}(f) = 1 \end{aligned} \tag{1-5}$$

其中，$\boldsymbol{\mathcal{R}}^{(i+n)}(f)$表示干扰信号 i 和噪声信号 n 的协方差矩阵，$\boldsymbol{r}(f)$为波束成形器指向目标声源信号的导向矢量。此外，自适应波束成形算法的设计准则还有最大输出信噪比(Maximum Output Signal to Noise Ratio, MSNR)准则，即通过最大化估计的目标语音信号与噪声信号的功率比值，求解最优加权系数，表示如下：

$$\max_{w} \frac{\boldsymbol{\mathcal{W}}^{\mathrm{H}}(f)\boldsymbol{\mathcal{R}}^{(s)}(f)\boldsymbol{\mathcal{W}}(f)}{\boldsymbol{\mathcal{W}}^{\mathrm{H}}(f)\boldsymbol{\mathcal{R}}^{(i+n)}(f)\boldsymbol{\mathcal{W}}(f)} \tag{1-6}$$

其中$\boldsymbol{\mathcal{R}}^{(s)}(f)$表示目标源信号 s 的空间协方差矩阵。

在真实场景中，$\boldsymbol{\mathcal{R}}^{(s)}(f)$和$\boldsymbol{\mathcal{R}}^{(i+n)}(f)$的获取往往非常困难，如在 MVDR 算法中由于无法知道具体的$\boldsymbol{\mathcal{R}}^{(i+n)}(f)$值，往往使用阵列接收到信号的空间协方差矩阵$\boldsymbol{\mathcal{R}}^{(x)}(f)$代替之，这种做法不仅限制了算法语音增强的性能上限，往往还造成算法稳健性的下降。一些对角加载[12]方案被设计用于缓解这类问题，但合适的对角加载值的选取仍然非常困难。Buckley 提出了线性约束最小方差(Linearly Constrained Minimum Variance, LCMV)[13]波束成形器对 MVDR 的问题进行了改进，LCMV 的思想是处理阵列信号时添加更多的陷零约束，从而达到更好的抑制非目

标声源的效果。文献[14]使用软定义的信号失真比共形表面阵列进一步提升了LCMV 的性能，但此算法往往导致计算资源的较大占用，且无法保证算法的实时性。

通过以上分析可知，基于波束成形的语音增强算法，往往需要知道麦克风阵列的准确结构特征，同时对阵列接收到信号的不同成分进行估计。在实际应用中，由于不可避免的阵列结构测量误差和声源定位误差，以及估计阵列信号空间协方差矩阵等信息时存在奇异值等问题，基于波束成形的语音增强算法往往无法达到理想的语音增强效果，在某些场景下甚至会出现严重的性能恶化。

1.2.2 基于盲源分离的语音增强算法

基于盲源分离的语音增强算法的核心思想是根据统计特性对声阵列接收到的信号直接进行目标和非目标信号分量分离，其主要处理步骤如图 1-3 所示。

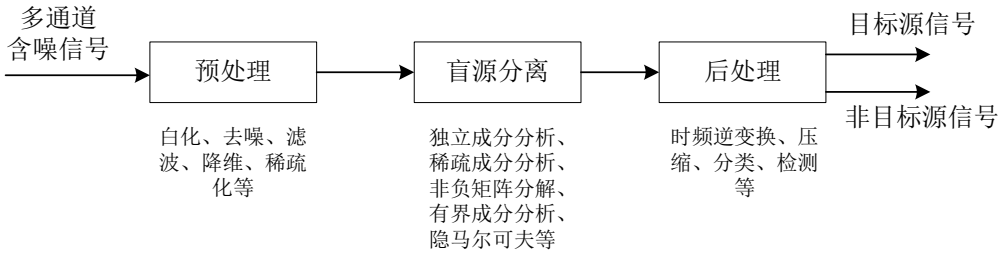


图 1-3 盲源分离算法主要步骤

Figure 1-3 The main steps of the blind source separation algorithm

基于盲源分离思想的语音增强算法可分为基于非负矩阵分解(Non-negative Matrix Factorization, NMF)[15]的盲源分离算法、基于隐马尔可夫模型(Hidden Markov Model, HMM)[16]的盲源分离算法等。基于 NMF 的盲源分离算法假设含噪语音信号中噪声与干净语音是相互独立的，含噪语音信号的幅度谱等特征往往具有非负特性，通过构建相互正交的干净语音信号的基向量和噪声信号基向量，将含噪语音幅度谱矩阵 M_{noisy} 分解为干净语音幅度谱矩阵 M_{clean} 和噪声信号幅度谱矩阵 M_{noise} 的乘积：

$$M_{noisy} = M_{clean} \times M_{noise} \tag{1-7}$$

其中，M_{noisy}、M_{clean} 和 M_{noise} 均为非负的矩阵。基于 NMF 的语音增强主要分为两个步骤：训练和增强。在训练阶段，利用干净语音数据和噪声数据获取干净语音特征基和噪声特征基；在增强阶段，将沿用训练阶段得到的干净语音特征基和噪声特征基组合成新的特征基，在含噪信号一致的情况下求解干净语音信号和噪声信号。此类方法假设噪声和干净语音相互独立，所以当噪声和语音特征接近时语

音增强效果往往较差。

基于 HMM 的盲源分离算法假设含噪语音、干净语音和噪声信号数据均满足一定的数学分布，如高斯分布等，同时假设模型中干净语音和噪声是相互独立的。根据语音序列的前后关联特性，对语音和噪声进行建模，计算某一时刻语音信号和噪声信号的存在概率。设干净语音从 0 到 t 时刻的概率密度函数为：

$$p\left(s_0^t\right)=\sum_{m_0^t}\prod_{\tau=0}^{t} b_{m_{\tau-1}m_\tau}\, p\left(s_\tau \mid m_\tau\right) \tag{1-8}$$

其中，m_0^t 表示从 0 时刻到 t 时刻的状态序列，$b_{m_{\tau-1}m_\tau}$ 表示从 $m_{\tau-1}$ 到 m_τ 的转移概率，$p\left(s_\tau \mid m_\tau\right)$ 表示 s_τ 对 m_τ 的条件概率。相应的，噪声信号和含噪信号的概率密度函数分别表示如下：

$$p\left(n_0^t\right)=\sum_{\bar{m}_0^t}\prod_{\tau=0}^{t} b_{\bar{m}_{\tau-1}\bar{m}_\tau}\, p\left(n_\tau \mid \bar{m}_\tau\right) \tag{1-9}$$

$$p\left(x_0^t\right)=\sum_{\tilde{m}_0^t}\prod_{\tau=0}^{t} b_{\tilde{m}_{\tau-1}\tilde{m}_\tau}\, p\left(s_\tau \mid \tilde{m}_\tau\right) \tag{1-10}$$

因此最终估计的目标语音信号为：

$$y_t=\mathrm{E}\left[s_t \mid x_0^{\mathrm{T}}\right]=\sum_{\tilde{m}_t} p\left(\tilde{m}_t \mid x_0^{\mathrm{T}}\right)\mathrm{E}\left[s_t \mid x_t, \tilde{y}_t\right] \tag{1-11}$$

由于基于 HMM 的语音增强对信号关系和数据分布都进行了假设，在进行语音增强时如果噪声或者语音的状态空间过大，会使得最优路径搜索变得困难，但状态空间过小又会使得模型对数据的拟合不足。

通过上述分析可知，基于盲源分离的语音增强算法不需要获取阵列结构、声源定位方向等先验信息[17]，而是根据信号的统计模型进行含噪信号的干净语音成分和噪声成分提取，然而在复杂场景中，语音信号分布往往不能用单一统计模型准确描述，同时干净语音信号和噪声信号也具有相互影响的关系，许多情况下不满足独立性假设，因此会限制基于盲源分离的语音增强算法的性能上限。

1.2.3 基于 DNN 的多通道语音增强算法

当前深度学习研究正蓬勃发展，各种有效的特征提取网络层出不穷，为多通道语音增强技术突破实际应用性能瓶颈、消除理论模型假设和先验误差限制、增强复杂场景适应性提供了重要思路。基于 DNN 的多通道语音增强算法以干净语音信号为目标对大量的含噪语音信号进行有监督训练，提取含噪信号中的干净语音分量特征，从而获得增强后的目标源语音。此类方法不需要阵列结构及声源定位等先验信息，不要求数据分布模型及独立性假设，近年来已经成为研究的主流[18]。

CNN 作为图像领域广泛使用的特征提取器也被引入到语音信号处理中[19]，将语音信号的短时傅里叶变换功率谱按照图像卷积的方式进行信号时频特征的提取，

从而建立含噪语音到纯净语音的映射关系。然而其往往需要很深的卷积层来进行高维全局特征的提取，并通过网络变形(如因果卷积[20]、膨胀卷积[21]等)来补充语音信号的时序关联特征和应对不定长输入。CRN 网络采用 CNN 和 RNN 相结合的思路(例如 CNN-LSTM-CNN 串接结构)，不仅能对语音信号频谱图进行时频特征的提取，也能对语音信号的时间序列特性进行建模，取得了很好的语音增强训练效果[22]。文献[9]将 GAN 用于语音增强领域，把含噪语音作为生成器的输入，鉴别器对生成器的输出和真实纯净语音进行鉴别，实验结果证明该模型在多项客观与主观指标上均优于传统方法的语音增强效果。文献[23]将 Transformer 中的注意力机制引入到语音增强中，通过计算输入信息的所有匹配位置的相似性得分来高度关注目标语音帧，对噪声语音帧分配更少的注意力，其语音增强性能优于 GAN 网络和 LSTM 网络。然而，由于这种方法需要评估所有样点间的相似性得分，其对于长输入任务会产生巨大的计算复杂度，其计算时间和存储空间需求呈平方增长，且层深以及参数的规模也限制了其在小型化智能语音交互系统的应用。为了适应实际应用中长程注意力的需求，许多兼具效率和速度的改进方法被提出，但更为通用的方法是基于稀疏注意力机制的算法，然而这些方法尚未在语音增强领域得到应用。

文献[24]首次提出了基于多层神经网络的双耳阵列语音分离，使用理想二值掩膜对语音信号和噪声信号进行估计，相比于传统方法中噪声源与目标源在同一方向时增强效果较差的问题，此方法能够取得更好的语音增强效果。文献[25]采用多层神经网络以幅度谱和耳间相位差为输入特征构建谱特征映射模型，结合自动语音识别相关指标的评估结果，该模型增强的语音有比较优异的语音增强效果。文献[26]设计了基于时域卷积的多通道信号去噪自动编码器进行语音增强，将时域的含噪波形信号直接映射为干净语音波形信号，相比于使用谱映射的方法，该方法不需要进行相位建模。文献[27]基于 UNet 结构设计了端到端的多通道语音增强系统，提取谱特征并设计了跨通道注意力模块提取网络信道特征，实现了在混响和嘈杂环境中的较高的语音增强性能且降低了语音识别的错误率。文献[28]提出了基于复频谱的多通道语音增强算法，以含噪信号频域实部和虚部为输入并进行复数卷积操作，提升了语音增强的性能上限，在有混响和无混响场景下均有较好的效果。文献[29]进行了多通道语音增强的图网络探索，将两层图卷积网络嵌入到UNet网络结构中提取信号的时间和空间信息以获得更好的语音增强效果，然而此方法直接使用了原始的图卷积网络结构，而没有考虑声学信号的性质及时空特征之间的关系，网络规模较大，因此尚无法应用于小型智能语音交互设备中。

1.3 多通道语音增强算法面临的挑战

尽管多通道语音增强算法已经取得了大量的研究进展，但是该领域仍然存在一些关键瓶颈问题亟待解决。

(1) 声阵列信号的空间关联性提取问题

声阵列信号中存在着复杂的空间关联性，受声源位置、阵列结构、声源速度、声学环境等多种条件的影响。然而，传统波束成形方法对多通道信号进行空域滤波，其受阵列结构误差、声源定位误差等影响，在实际应用中无法达到理论最优的性能。而现有的深度学习方法仅简单挪用其他领域的机器学习网络结构，没有结合声阵列信号的独特机理来进行网络模型的高效性、可控性和可解释性改造，尚无法适用于小型智能人机交互设备。此外，由于缺乏准确的阵列分布先验信息，难以对多通道信号间的空间关联关系进行完全解析，导致实际应用中的声源盲分离效果不佳。

(2) 时-空依赖性融合问题

声阵列信号中存在着的空间关联性特征和时频关联性特征，且空间信息和时频信息相互影响相互关联。当前的语音增强算法缺乏对声阵列信号的复杂时空频关联性的分析和建模，未能有效利用声阵列信号的时-空依赖关系进行多源分离和目标语音重建。

(3) 目标语音重建的相位信息损失问题

当前以幅度谱等具有清晰谐波结构的谱特征为输入信号进行处理，重建目标语音时，往往直接使用含噪信号的相位谱特征，使得估计的目标语音信号携带了相位谱中的噪声信息，如何有效利用输入语音信号的全部信息进行语音目标增强是进一步提升实际应用系统性能的关键。

(4) 语义信息特征感知问题

网络训练过程中对输入的含噪信号进行目标与非目标特征的提取，但很少考虑人语言表达的语义信息，比如语音音素必然具有一定的发音时长的关联，因此需要针对噪声与语音之间的语义差异特征进行噪声感知或语音感知。

(5) 损失函数与人听觉感知不匹配问题

传统深度学习网络训练的损失函数往往基于信号的对比指标，而不是基于人对不同频段信号的听觉感知差异，难以匹配不同频段的语音可感知度，因此设计符合人听觉感知的网络训练损失函数对于实现智能交互设备中的语音增强质量的提升有着较为重要的意义。

1.4 研究目标与内容

本文针对阵列结构未知的陌生多变场景，挖掘多通道混合信号中所包含的多源信号时变特征和阵列空间相关性特征，设计构建基于时空图卷积语音增强网络的实时高效的多通道语音增强算法，提升目标语音增强效果和鲁棒性，并针对系统的实时性和实际应用瓶颈问题进行了优化设计，进一步为算法在小型智能人机交互设备的落地应用提供可行方案和技术支持。

研究内容包括：

(1) 在缺失阵列和场景先验信息的情况下，基于非欧空间图理论，进行多通道语音增强问题的图建模，解析声阵列信号中所隐含的与阵列拓扑和声源位置相关的空间关联关系。具体包括：针对麦克风时空数据的图结构建模、多通道时空图聚合运算设计、可反映长时通道关联关系的动态自适应邻接矩阵构建和图神经网络搭建。

(2) 融合基于图卷积运算的空间关联性提取方法和基于时频卷积的时域关联性提取方法，构建时空图卷积语音增强网络(Spatial Temporal Graph Convolution Speech Enhancement Network, STGCSEN)，提取声阵列信号的空间关联性、时间关联性和时-空依赖性特征，在缺失场景和阵列先验信息的情况下，显著提升目标语音重建质量，抑制噪声和干扰。实验证明，本文提出的 STGCSEN 在多种恶劣场景下的主客观指标均达到了最优或接近最优效果。

(3) 为了进一步解决语音增强网络应用于小型智能人机交互设备所面临的瓶颈，针对目标语音重建的相位信息损失问题，进行了基于复频谱的时空图卷积语音增强网络拓展；进行了网络参数量和系统实时性的优化；进行了基于人耳听觉感知的语音可懂度网络训练损失函数设计。实验证明，基于复频谱的时空图卷积语音增强网络提升了目标语音增强性能的上限，网络参数量和实时性优化为算法的工程落地提供了技术支持，基于语音可懂度的损失函数可使网络输出的重建目标语音更符合人耳听觉感知。

1.5 论文结构安排

本文的第一章根据多通道语音增强的实际应用场景需要，引出了该问题的研究背景及意义。接着针对多通道语音增强问题的研究现状进行归纳梳理，详细介绍了基于波束成形的语音增强算法、基于盲源分离的多通道语音增强算法和基于 DNN 的多通道语音增强算法，分析了各类算法的核心思想及优缺点，并针对当前语音增强领域存的问题提出了本文的研究目标与内容。

第二章完成了多通道语音增强的图理论建模和设计。首先对空间信号传播过程和阵列信号处理过程进行建模，接着根据建模结果进行基于麦克风阵列的图构建和声阵列信号的图卷积聚合操作，根据阵列不同阵元接收信号的时延及衰减差异完成了图上邻接矩阵的构建，最终完成了基于声阵列信号的图神经网络搭建。

第三章构建了时空图卷积语音增强网络。首先根据图卷积聚合操作设计了空间信息提取模块，对声阵列信号的空间关联性进行提取；接着设计了基于二维卷积运算的时频信息提取模块，实现了基于时间关联性的时频信息提取；然后根据时间关联性和空间关联性的相互影响相互依赖关系设计了时-空依赖性融合模组(Spatial-Temporal Module, ST)；设计了通道信息融合模块，将级联的 ST 模组输出的多通道增强语音信号融合为增强的单通道语音信号输出；最后，进行了多算法对比实验、主观语音质量评价实验、多邻接矩阵对比实验和多数据集对比试验，实验结果证明了与现有算法相比，本文所设计算法在多噪声场景、多声源类型均取得了最优或较优的效果。

第四章提出了时空图卷积网络的优化方案。首先分析了以幅度谱作为输入特征时的相位信息损失问题，提出了基于复数相乘原理的时空图卷积语音增强网络的复频谱拓展方法，实验结果证明了复频谱时空图卷积语音增强能够取得更好的语音增强效果，提升了语音增强的性能上限；接着进行了网络参数量-语音增强性能-算法耗时的比较实验，给出了不同场景下应选择的最优网络结构，为算法的实际工程落地提供了理论依据；最后设计了基于语音可懂度指数(Speech Intelligibility Index, SII)的损失函数，使得网络在训练过程中能够对人耳感知语音更明显的频带分配更多的权重，实验证明基于 SII 的损失函数可使网络输出的语音更符合人耳的听觉感知。

第五章对本文的研究工作进行总结，分析了本文工作的不足并对未来工作进行了展望。

2 声阵列信号的图理论建模与设计

现有的基于信号模型的语音增强算法由于缺乏准确的阵列分布先验信息，难以对多通道信号间的空间关联关系进行完全解析，导致实际应用中的声源盲分离效果不佳；而基于机器学习的语音增强算法仅仅简单挪用其他领域的网络结构，没有结合声阵列信号的独特机理来进行网络模型的高效性、可控性和可解释性改造，尚无法适用于小型智能人机交互设备。这些方法都缺少对多通道声阵列信号的空间关联性、时频关联性和时空依赖性的分析和建模，从而加剧语音增强过程中的目标与非目标信号分量的混叠。因此，本章首先根据智能语音交互三维环境下的声阵列信号传播模型完成多通道语音增强问题的建模；其次，针对麦克风阵列信号的独特机理进行节点与图理论的构建和图聚合运算的设计；然后，设计动态自适应邻接矩阵以获取节点间特征信息提取多通道信号空间相关性；最后，构建了基于多通道信号语音增强场景的图神经网络，为后续设计时空图卷积语音增强网络模型奠定理论基础。

2.1 多通道语音增强理论建模

在三维空间多声源环境中，使用麦克风阵列对目标声源、干扰源和噪声进行声信号录制，麦克风阵列除了接收到目标语音外，还会接收到干扰人声、混响和机器噪声等多种音频信号。多通道语音增强算法对麦克风阵列接收到的混合音频信号进行处理，通过估计每个通道的信号加权系数从混合信号中恢复目标语音信号，同时抑制干扰及噪声信号。

设目标声源信号为$s(t;\boldsymbol{r}_s)$，$\boldsymbol{r}_s$表示目标声源位置。假设有K个点噪声源，位于$\boldsymbol{r}_k$的点噪声信号为$n(t;\boldsymbol{r}_k)$，背景白噪为$n_0(t)$，麦克风阵列共C个麦克风，$\boldsymbol{r}_c$表示第c个麦克风的位置。因此，第c个麦克风接收到的混合含噪信号的时域表达式为：

$$x(t;\boldsymbol{r}_s,\boldsymbol{r}_c)=h(t;\boldsymbol{r}_s,\boldsymbol{r}_c)*s(t;\boldsymbol{r}_s)+\sum_{k=1}^{\mathrm{K}}h(t;\boldsymbol{r}_k,\boldsymbol{r}_c)*n(t;\boldsymbol{r}_k)+n_0(t) \tag{2-1}$$

其中，$*$表示卷积运算符，$h(t;\boldsymbol{r}_s,\boldsymbol{r}_c)$表示声波从声源位置$\boldsymbol{r}_s$到第$c$个麦克风位置$\boldsymbol{r}_c$的传递函数。当不考虑空间混响时：

$$h(t;\boldsymbol{r}_s,\boldsymbol{r}_c)=\boldsymbol{\alpha}(|\boldsymbol{r}_s-\boldsymbol{r}_c|)\delta\left(t-\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|}{\mu}\right) \tag{2-2}$$

其中，$\boldsymbol{\alpha}(|\boldsymbol{r}_s-\boldsymbol{r}_c|)$表示声波传播的幅度衰减函数，$\delta\left(t-\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|}{\mu}\right)$表示声波的传播时

延，μ 表示声音传播速度。因此，第 c 个麦克风接收到的混合含噪信号的频域表达式为：

$$\begin{aligned} X(f;\boldsymbol{r}_s,\boldsymbol{r}_c) &= \mathcal{A}(f,|\boldsymbol{r}_s-\boldsymbol{r}_c|)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|}{\mu}\right)S(f;\boldsymbol{r}_s) \\ &\quad+\sum_{k=1}^{\mathrm{K}}\mathcal{A}(f,|\boldsymbol{r}_k-\boldsymbol{r}_c|)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_k-\boldsymbol{r}_c|}{\mu}\right)N(f;\boldsymbol{r}_k)+N_0(f) \end{aligned} \tag{2-3}$$

其中，$\mathcal{A}(f,|\boldsymbol{r}_s-\boldsymbol{r}_c|)=\exp\left(-\frac{\xi f^2}{\mu^3}|\boldsymbol{r}_s-\boldsymbol{r}_c|\right)$，$\xi$ 是常数，与介质密度、声音传播速度、粘滞系数、热传导系数、等容热容和等压热容有关[30]，当考虑在 $20^{\circ}\mathrm{C}$ 干燥的室内环境中一般取 5.58×10^{-4}，具体推导过程见附录 A。

构建波束成形器，从麦克风阵列接收到的多通道含噪信号中恢复出目标语音信号。令波束成形器的焦点(Focal Point)为 $\boldsymbol{r}_i$，波束成形器输出的目标信号估计为：

$$\begin{aligned} \hat{S}(f;\boldsymbol{r}_i) &= \sum_{c=1}^{\mathrm{C}}\mathcal{B}_{ic}X(f;\boldsymbol{r}_s,\boldsymbol{r}_c) \\ &=\sum_{c=1}^{\mathrm{C}}\mathcal{B}_{ic}\mathcal{A}(f,|\boldsymbol{r}_s-\boldsymbol{r}_c|)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|}{\mu}\right)S(f;\boldsymbol{r}_s) \\ &\quad+\sum_{c=1}^{\mathrm{C}}\sum_{k=1}^{\mathrm{K}}\mathcal{B}_{ic}\mathcal{A}(f,|\boldsymbol{r}_k-\boldsymbol{r}_c|)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_k-\boldsymbol{r}_c|}{\mu}\right)N(f;\boldsymbol{r}_k)+\sum_{c=1}^{\mathrm{C}}\mathcal{B}_{ic}N_0(f) \end{aligned} \tag{2-4}$$

其中 $\mathcal{B}_{ic}$ 表示波束成形器焦点 $\boldsymbol{r}_i$ 到第 c 个麦克风多通道加权系数。此时输出信号的信噪比(Signal-to-Noise Ratio, SNR)可以表示为：

$$SNR(\boldsymbol{r}_s,\boldsymbol{r}_i)=\frac{\left\|\sum_{c=1}^{\mathrm{C}}\mathcal{B}_{ic}\mathcal{A}(f,|\boldsymbol{r}_s-\boldsymbol{r}_c|)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|}{\mu}\right)S(f;\boldsymbol{r}_s)\right\|^2}{\left\|\sum_{c=1}^{\mathrm{C}}\sum_{k=1}^{\mathrm{K}}\mathcal{B}_{ic}\mathcal{A}(f,|\boldsymbol{r}_k-\boldsymbol{r}_c|)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_k-\boldsymbol{r}_c|}{\mu}\right)N(f;\boldsymbol{r}_k)+\sum_{c=1}^{\mathrm{C}}\mathcal{B}_{ic}N_0(f)\right\|^2} \tag{2-5}$$

其中，$\|\bullet\|^2$ 表示全频谱的信号的功率，因此波束形成器的设计则可以表示为求解最优的多通道加权系数集 $\mathcal{B}$ 的问题，即：

$$\mathcal{B}^{(\mathrm{optimal})}=\arg\max_{\mathcal{B}}\left\langle\mathop{\mathrm{E}}_{\boldsymbol{r}_i\in\mathrm{FOV}}\left[SNR(\boldsymbol{r}_s,\boldsymbol{r}_i)\right]\right\rangle \tag{2-6}$$

其中，FOV (Field of View)代表关注区域，即目标与噪声声源可能出现的源空间。

由公式(2-5)和公式(2-6)可知，当不考虑定位误差时，理想的波束成形器沿目标声源方向($\boldsymbol{r}_i=\boldsymbol{r}_s$)的增益最大，沿非目标源方向($\boldsymbol{r}_i\neq\boldsymbol{r}_s$)形成零陷，从而得到最优的目标重建信噪比。下面将就最常见的延迟-求和波束成形器(Delay and Sum Beamformer, DSB)和最小方差无失真响应波束成形器(MVDR)的求解给出理论分析。

(1) DSB 波束成形器

DSB 根据各源信号到达阵列中每个麦克风的时间不同，选定其中一个通道信

号为参考信号，比较信号到达其他麦克风相对参考信号的时延，根据信号传播过程中的时间延迟和幅度衰减公式，对阵列信号进行幅值和时延的补偿，即令：

$$\begin{aligned}\boldsymbol{\mathcal{B}}_{ic} &= \mathcal{A}^{-1}\left(f,|\boldsymbol{r}_i-\boldsymbol{r}_c|\right)\exp\left(j2\pi f\frac{|\boldsymbol{r}_i-\boldsymbol{r}_c|}{\mu}\right)\\ &= \exp\left(\frac{\xi f^2}{\mu^3}|\boldsymbol{r}_i-\boldsymbol{r}_c|\right)\exp\left(j2\pi f\frac{|\boldsymbol{r}_i-\boldsymbol{r}_c|}{\mu}\right)\end{aligned} \tag{2-7}$$

可得估计的目标信号：

$$\begin{aligned}\hat{S}(f;\boldsymbol{r}_i) = &\sum_{c=1}^{\mathrm{C}}\exp\left(\frac{\xi f^2}{\mu^3}|\boldsymbol{r}_i-\boldsymbol{r}_c|\right)\exp\left(j2\pi f\frac{|\boldsymbol{r}_i-\boldsymbol{r}_c|}{\mu}\right)\\ &\exp\left(-\frac{\xi f^2}{\mu^3}|\boldsymbol{r}_s-\boldsymbol{r}_c|\right)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|}{\mu}\right)S(f;\boldsymbol{r}_s)\\ &+\sum_{c=1}^{\mathrm{C}}\sum_{k=1}^{\mathrm{K}}\exp\left(\frac{\xi f^2}{\mu^3}|\boldsymbol{r}_i-\boldsymbol{r}_c|\right)\exp\left(j2\pi f\frac{|\boldsymbol{r}_i-\boldsymbol{r}_c|}{\mu}\right)\\ &\exp\left(-\frac{\xi f^2}{\mu^3}|\boldsymbol{r}_k-\boldsymbol{r}_c|\right)\exp\left(-j2\pi f\frac{|\boldsymbol{r}_k-\boldsymbol{r}_c|}{\mu}\right)N(f;\boldsymbol{r}_k)\\ &+\sum_{c=1}^{\mathrm{C}}\boldsymbol{\mathcal{B}}_{ic}N_0(f)\end{aligned} \tag{2-8}$$

当忽略声源定位误差，波束成形器焦点与声源位置相同，即 $\boldsymbol{r}_i=\boldsymbol{r}_s$ 时，估计的目标信号可表示为：

$$\begin{aligned}\hat{S}(f;\boldsymbol{r}_i) = &\sum_{c=1}^{\mathrm{C}}S(f;\boldsymbol{r}_s)\\ &+\sum_{c=1}^{\mathrm{C}}\sum_{k=1}^{\mathrm{K}}\exp\left(\left(\frac{\xi f^2}{\mu^2}+j2\pi f\right)\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|-|\boldsymbol{r}_k-\boldsymbol{r}_c|}{\mu}\right)\\ &+N_0(f)\end{aligned} \tag{2-9}$$

当 $\boldsymbol{r}_i=\boldsymbol{r}_s$ 时，$\boldsymbol{\mathcal{B}}_{ic}=\boldsymbol{\mathcal{B}}_{sc}=\exp\left(\frac{\xi f^2}{\mu^3}|\boldsymbol{r}_s-\boldsymbol{r}_c|\right)\exp\left(j2\pi f\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|}{\mu}\right)$，该阵列信号的SNR表达式为：

$$\mathrm{SNR}=\frac{S(f;\boldsymbol{r}_s)}{\sum_{k=1}^{\mathrm{K}}\mathrm{E}\left[\exp\left(\left(\frac{\xi f^2}{\mu^2}+j2\pi f\right)\frac{|\boldsymbol{r}_s-\boldsymbol{r}_c|-|\boldsymbol{r}_k-\boldsymbol{r}_c|}{\mu}\right)\right]N(f;\boldsymbol{r}_k)} \tag{2-10}$$

在波束成形语音增强过程中，我们希望SNR尽可能大，因此波束成形器在对阵列信号进行处理时，应当增加对目标声源信号的时延和衰减补偿的相干水平，增加对噪声源信号的时延和衰减的非相干水平，以对噪声分量进行去相关。上式中，$\boldsymbol{r}_s$ 处的目标源信号通过对信号时延和衰减的相干相加得到加强，分母可以看成C个复平面的向量，第 c 个麦克风的向量角为 $\frac{2\pi}{\mu}\left(|\boldsymbol{r}_s-\boldsymbol{r}_c|-|\boldsymbol{r}_k-\boldsymbol{r}_c|\right)$，当麦克风的差分

路径距离(Differential Path Distance, DPD)分布的 Pielou 均匀度指数(Pielou's Evenness Index, PEI)足够大时会使得各通道信号进行不相干叠加[31]，即使得分母中对噪声的功率增益接近于零，即 $\mathrm{E}\left(\left|\boldsymbol{r}_s-\boldsymbol{r}_c\right|-\left|\boldsymbol{r}_k-\boldsymbol{r}_c\right|\right)\to 0$，因此波束形成器输出 SNR 达到最高，取得最好的语音增强效果。即当 $\boldsymbol{r}_i=\boldsymbol{r}_s$ 时得到最优的多通道加权系数：

$$\boldsymbol{\mathcal{B}}_{ic}=\boldsymbol{\mathcal{B}}_{sc}=\exp\left(\frac{\xi f^2}{\mu^3}\left|\boldsymbol{r}_s-\boldsymbol{r}_c\right|\right)\exp\left(j2\pi f\frac{\left|\boldsymbol{r}_s-\boldsymbol{r}_c\right|}{\mu}\right) \tag{2-11}$$

此时该波束成形器取得最好的语音增强效果，估计的目标源信号为：

$$\hat{S}\left(f;\boldsymbol{r}_s\right)=\sum_{c=1}^{\mathrm{C}}\exp\left(\frac{\xi f^2}{\mu^3}\left|\boldsymbol{r}_s-\boldsymbol{r}_c\right|\right)\exp\left(j2\pi f\frac{\left|\boldsymbol{r}_s-\boldsymbol{r}_c\right|}{\mu}\right)X\left(f;\boldsymbol{r}_s,\boldsymbol{r}_c\right) \tag{2-12}$$

(2) MVDR 波束成形器

MVDR 的基本思想是波束成形器沿目标声源的方向增益为 1，同时对干扰源方向形成零陷，使得系统输出总能量最小，公式如下：

$$\arg\min\left\langle\left\|X\left(f;\boldsymbol{r}_s,\boldsymbol{r}_c\right)\right\|^2\right\rangle\quad\text{s.t.}\ \sum_{c=1}^{\mathrm{C}}\boldsymbol{\mathcal{B}}_{ic}H\left(f;\boldsymbol{r}_i,\boldsymbol{r}_c\right)=1 \tag{2-13}$$

其中，$H\left(f;\boldsymbol{r}_i,\boldsymbol{r}_c\right)$ 为 $h\left(t;\boldsymbol{r}_i,\boldsymbol{r}_c\right)$ 的频域变换，则此时 MVDR 的输出为：

$$\begin{aligned}\hat{S}\left(f;\boldsymbol{r}_i\right)=&\sum_{c=1}^{\mathrm{C}}\boldsymbol{\mathcal{B}}_{ic}H\left(f;\boldsymbol{r}_i,\boldsymbol{r}_c\right)S\left(f;\boldsymbol{r}_s\right)\\&+\sum_{c=1}^{\mathrm{C}}\sum_{k=1}^{\mathrm{K}}\boldsymbol{\mathcal{B}}_{ic}H\left(f;\boldsymbol{r}_k,\boldsymbol{r}_c\right)N\left(f;\boldsymbol{r}_k\right)+N_0\left(f\right)\end{aligned} \tag{2-14}$$

则 MVDR 的理论目标函数为：

$$\boldsymbol{\mathcal{B}}_{\mathrm{MVDR}}^{(\mathrm{optimal})}=\arg\min_{\boldsymbol{\mathcal{B}}}\left\langle\sum_{c=1}^{\mathrm{C}}\sum_{k=1}^{\mathrm{K}}\boldsymbol{\mathcal{B}}_{ic}H\left(f;\boldsymbol{r}_k,\boldsymbol{r}_c\right)N\left(f;\boldsymbol{r}_k\right)\right\rangle \tag{2-15}$$

根据公式(2-9)、(2-11)和(2-15)的分析可知，波束成形器在恢复目标源信号时，首先在每个频点处对多通道信号的幅度和相位进行调整，然后对其求和得到估计的目标源信号。

综上，多通道语音增强算法根据阵列结构和声信号传播模型为每个通道赋予合适的加权系数，进行声阵列信号时频点特征的通道间聚合，从而增强目标源信号分量并抑制干扰噪声分量。从公式(2-4)可以看出加权系数与麦克风的位置和声源的位置有关，而在实际场景中，声源定位和麦克风位置测量往往存在不可避免的误差，这会导致波束成形器的语音增强性能严重下降。为此，本文拟采用基于自适应时空图卷积的方法来聚合多通道信号，通过训练使网络在缺少先验知识或先验知识存在误差的情况下实现较好的语音增强性能。

2.2 声阵列信号的图构建

从 2.1 节多通道语音增强建模过程可以发现，阵列接收到的多通道信号的振幅和相位差可以反映与分布式麦克风位置相关的目标语音分量的空间依赖性。因此，本节将每个麦克风采集到的时频数据作为分布在非欧几里得空间中的图节点，基于与声源位置、阵列拓扑、声学环境强相关的图结构来聚合邻居节点的信息，构建图卷积网络(Graph Convolution Network, GCN)来动态捕捉多通道含噪语音的空间关联性，实现目标语音重建。本文提出的这种方法不需要阵列和声学场景等先验信息，从而避免了传统波束成形算法中由于麦克风位置测量不准确等产生的误差而导致算法性能恶化的问题。另外由于图运算不需要深层网络，从而非常适合在小型智能交互设备中应用。

2.2.1 节点与多通道信号图构建

基于声阵列信号机理，按照非欧几里得空间节点特点，将各麦克风通道构建为图上的节点，阵列信号时频特征构建为图上节点特征，利用图结构的边权重来表示阵列信号通道间的关联关系。

针对声阵列信号的图结构可以表示为$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，其中$\boldsymbol{V}$为图节点集合，且$|\boldsymbol{V}|=$ C，即阵列内有C 个麦克风。$\boldsymbol{E}$为边的集合表示两个麦克风接收到的信号之间的连接关系。邻接矩阵$\boldsymbol{A}=(a_{ij})_{C\times C}\in\mathbb{R}^{C\times C}$表示两个麦克风通道信号之间连接关系的权重，表征两个麦克风接收到的信号之间的空间相关性，其中a_{ij}为第 j 个麦克风接收到信号到第i个麦克风接收到信号的权重。度矩阵$\boldsymbol{D}\in\mathbb{R}^{C\times C}$是一个对角矩阵且$\boldsymbol{D}_{ii}\in\sum_j a_{ij}$。

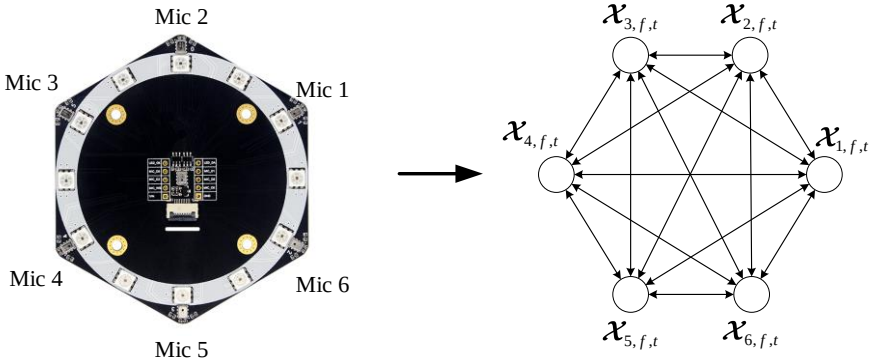


图 2-1 麦克风阵列节点与图构建

Figure 2-1 Microphone array node and graph construction

由 2.1 节定义，麦克风阵列的多通道含噪语音信号的幅度谱表示为$\boldsymbol{\mathcal{X}}\in\mathbb{R}^{C\times F\times T}$，

幅度谱可以较为清晰的展示语音的谐波结构，如图 2-2 所示。因此，本文采用幅度谱特征为模型输入，对C 个通道的幅度谱进行多通道语音增强得到估计的目标语音幅度谱为$\hat{\boldsymbol{\mathcal{X}}} \in \mathbb{R}^{F\times T}$，则图上的多通道语音增强任务可以描述如下：

$$\hat{\boldsymbol{\mathcal{X}}} = \mathcal{G}\left(\boldsymbol{\mathcal{X}};\boldsymbol{G}\right) \tag{2-16}$$

$\mathcal{G}(\bullet)$表示在图上由幅度谱特征到估计的幅度谱之间的映射关系。

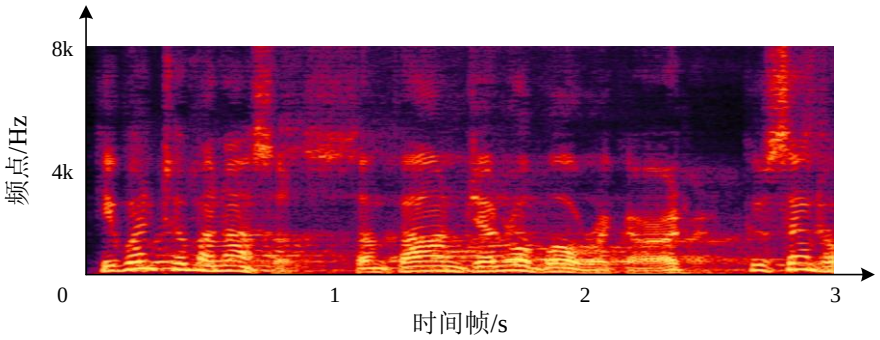


图 2-2 语音信号幅度谱

Figure 2-2 Speech Signal Amplitude Spectrum

因此，本节进行了麦克风阵列信号的图构建，将麦克风、阵列信号特征分别构建为图上节点和图结构，将阵列中通道信号的相关性构建为图上节点间权重，完成了将阵列信号从欧氏空间到非欧空间的理论建模，为后续阵列信号的图聚合运算、图神经网络搭建奠定了基础。

2.2.2 图聚合运算设计

假设第c个麦克风通道第 f 个频点和第t个时间帧的幅度谱特征为$\boldsymbol{\mathcal{X}}_{c,f,t}$，模拟麦克风阵列信号波束成形，由公式(2-4)可知增强目标源信号并抑制干扰噪声需要对通道间特征进行聚合。语音增强的图聚合过程由两步产生：(1) 聚合与节点c直接相邻通道的多时频点的关联特征；(2) 本节点的时频特征的传递。图聚合过程是在节点间发生的，因此图聚合可以被认为是相邻节点间的“消息传递”，节点特征每一次聚合都会从邻居节点那里接收到信息并更新自身特征，通过迭代的重复K 次图聚合过程，聚合的感受野可以有效扩大，从而获得图上全部节点信息。

对阵列信号第c个通道第t个时间帧幅度谱特征$\boldsymbol{\mathcal{X}}_{c,f,t}$沿邻居节点及自身节点进行一次图聚合表达如下：

$$\hat{\boldsymbol{\mathcal{X}}}_{c,f,t} = g\left(\frac{\sum_{i\in\boldsymbol{\mathcal{N}}(c)}\boldsymbol{W}_{i,f-\Delta f:f+\Delta f,t}\boldsymbol{\mathcal{X}}_{i,f-\Delta f:f+\Delta f,t}}{\left\|\boldsymbol{\mathcal{N}}\left(c\right)\right\|}+\boldsymbol{B}_{c,f,t}\boldsymbol{\mathcal{X}}_{c,f,t}\right) \tag{2-17}$$

其中，$\boldsymbol{\mathcal{N}}\left(c\right)$表示节点$c$的邻居节点的集合，$\boldsymbol{W}_{i,f-\Delta f:f+\Delta f,t}$表示$i$个邻居节点到$c$节

点的权重，$\boldsymbol{B}_{c,f,t}$表示节点c的自相关权重，g表示可训练的非线性函数，聚合过程如图 2-3 所示。

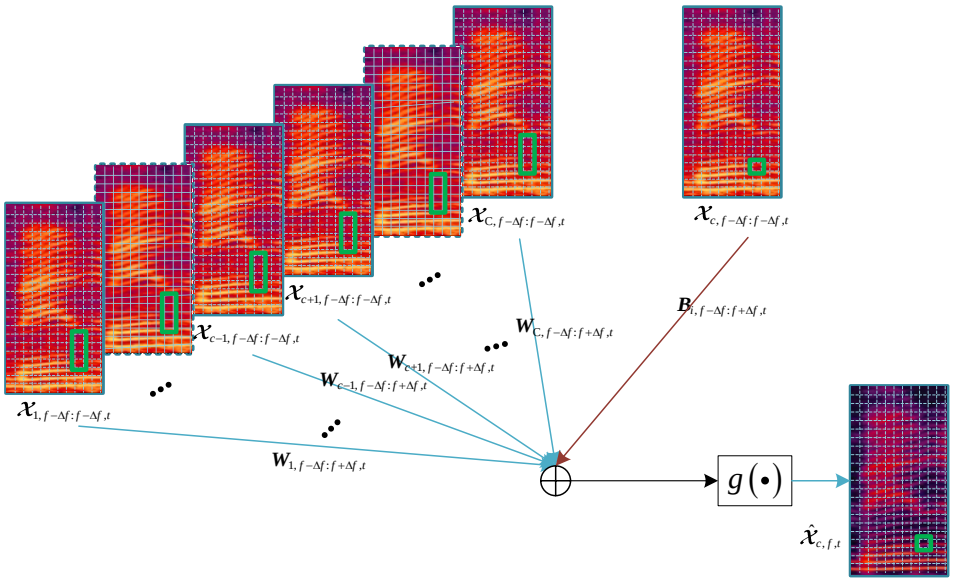


图 2-3 对第c个麦克风的第(f, t)时频点特征进行一次图聚合示意图

Figure 2-3 Schematic diagram of a graph aggregation of the (f, t) bins features of the c-th microphone

针对智能交互设备的典型应用场景，分别对多通道声阵列信号沿时域的关联性和频域的关联性进行讨论，从而确定信号进行时频域图聚合的合理范围。本文的多通道声阵列信号的采样率均为 16k Hz，对时域信号均进行 512 点(即帧长 $512/16000=32\text{ms}$)FFT 变换，每个频点覆盖了 31.25Hz 的频带宽度。当考虑相邻时间帧的关联性时，按照阵列间距最大 19 cm 及声速 340 m/s 进行计算，声波到达两个阵元时间差最大为 0.56 ms，远远小于 32/2=16 ms，所以第t个时间帧信号与其前一时间帧和后一时间帧在信息重合性上关联不大，因此不需要考虑图聚合运算中跨时间帧的聚合；当进行相邻频点的关联性分析时，考虑目标人可能正在移动，根据声音传递函数的频移特性及多普勒频移函数进行计算：

$$\Delta f=\frac{v}{\mu-v}f \tag{2-18}$$

其中，v为声源移动速度，μ为声音传播速度，f为声音信号频率。当声源在走路时，速度约为$v=2\text{m/s}$时，则 5281 Hz 以上的声信号相邻两个频点之间开始有关联，针对声源最高频率 20k Hz 时关联的频点数最多为 4 个；当声源在跑步时，速度约为$v=10\text{m/s}$，1031 Hz 以上的声信号，两个频点开始有关联，针对声源最高频率 20k Hz 时关联的频点数最多为 20 个。综合上述分析，在对阵列信号幅度谱特征进行图聚合时，沿时间维度可以对每一个时间帧的信号分别进行聚合，沿频率维度需要把每个时间帧上的相邻多个频点作为特征向量进行聚合，即需要构

造一个沿通道域和频点域的二维图聚合运算。

本节详细阐述了声阵列图聚合运算的设计思路，分析了多通道阵列信号在同一时刻相邻麦克风通道，相邻时间帧和相邻频点的信息关联性，得出需构造沿通道域和频点域的二维聚合方式的结论，并给出了聚合过程的数学表达和原理设计。

2.3 声阵列信号的邻接矩阵构建

声阵列信号的空间关联性融合了阵元间距离衰减函数、声波传播模型、信号时频变化模型和声源运动行为模式等多重因素，而声阵列信号图模型中的邻接矩阵是解析信号通道间空间关联性的关键。本节从理论上分析了图邻接矩阵与阵列信号通道间相关性的对应关系。针对波束成形中权重存在误差而导致算法出现性能恶化的问题，根据阵列间距、信号传播时延设计了不同类型的图上邻接矩阵，以更好地获得阵列信号目标语音分量的空间依赖性的特征。

由公式(2-7)可得到源信号到麦克风时的幅度衰减和相位差，以麦克风阵列第1个麦克风为参考麦克风，则第c个麦克风到参考麦克风的多通道加权系数表达如下：

$$\mathcal{B}_{1c}=\exp\left(\frac{\xi f^{2}}{\mu^{3}}\left|\boldsymbol{r}_{1}-\boldsymbol{r}_{c}\right|\right)\exp\left(j2\pi f\frac{\left|\boldsymbol{r}_{1}-\boldsymbol{r}_{c}\right|}{\mu}\right) \tag{2-19}$$

由于现实中无法对阵列中麦克风位置$\boldsymbol{r}_{c}$、阵列中麦克风间距及声速μ等影响因子进行准确地测量，所以无法得到准确的理论权重，对于整个图结构，其节点权重的集合由邻接矩阵$\boldsymbol{A}$表示，则对于t时刻整个图信号的聚合运算可表如下：

$$\hat{\boldsymbol{\mathcal{X}}}_{:,:,t}=\boldsymbol{g}\left(\boldsymbol{A}\boldsymbol{\mathcal{X}}_{:,:,t}\right) \tag{2-20}$$

因此在图上对阵列信号进行处理时，阵列信号间的权重由邻接矩阵$\boldsymbol{A}$表示，为使得邻接矩阵$\boldsymbol{A}$能够更有效的表达阵列信号通道间权重，我们依据理论关系设定初值，然后根据实际数据用图卷积网络对邻接矩阵进行学习和训练。

从不同的阵列类型、不同的应用场景下对节点关系进行不同的假设，可以设计不同的邻接矩阵结构和初值设置方式。本文所设计的邻接矩阵可分为固定邻接矩阵、动态邻接矩阵和进化邻接矩阵三类。

(1) 固定邻接矩阵(Fixed Adjacency Matrix)

如果节点间的相关性不随时间变化，可以根据节点间相关性设计固定邻接矩阵并在整个实验训练和测试过程中一直保持不变。固定邻接矩阵包括预定义的邻接矩阵各元素值或通过预训练获得的优化的邻接矩阵各元素值。拟采用的固定邻接矩阵类型为：反对称矩阵、实矩阵和复矩阵，固定矩阵各元素反映节点间的相关性，以距离衰减值、指数距离衰减值为元素值以设计固定邻接矩阵。

全 1 固定邻接矩阵：假设麦克风阵列的各个阵元两两之间互相影响，且影响权重视为 1，即对于任意的$i, j \in [1, C]$，邻接矩阵值如下：

$$a_{ij} = 1 \tag{2-21}$$

距离衰减邻接矩阵：声音信号在传播过程中随着距离而不断衰减，麦克风阵列中不同阵元位置不同，声源到达不同阵元引起的衰减也不同，我们以阵元间距的倒数表示这种衰减，即对于任意的$i, j \in [1, C]$且$i \neq j$，邻接矩阵值如下：

$$a_{ij} = \begin{cases} \dfrac{1}{d_{ij}} & i \neq j \\ 0 & i = j \end{cases} \tag{2-22}$$

其中d_{ij}表示阵元 j 到阵元i的距离。

指数距离衰减矩阵：声音信号在介质中传播过程中会出现散射衰减和吸收衰减，衰减程度与距离呈指数关系，因此指数距离衰减矩阵表示如下：

$$a_{ij} = \begin{cases} \exp\left(-\dfrac{1}{d_{ij}}\right) & i \neq j \\ 0 & i = j \end{cases} \tag{2-23}$$

反对称指数衰减矩阵：声音信号传递到不同阵元的时间有先后，假设节点i相对于节点 j 更接近声源信号，对于节点i而言，其受到节点 j 的权重为a_{ij}，对于节点 j 而言，其受到节点i的权重应为$1/a_{ij}$，因此反对称指数衰减矩阵表示如下：

$$a_{ij} = \begin{cases} \exp\left(-\dfrac{1}{d_{ij}}\right) & i > j \\ 0 & i = j \\ \exp\left(\dfrac{1}{d_{ij}}\right) & i < j \end{cases} \tag{2-24}$$

(2) 动态邻接矩阵

由于先验知识的缺陷或数据不完整，预定义的固定邻接矩阵并不能准确反映节点之间的真实依赖关系，因此本文拟利用动态邻接矩阵学习阵列信号间的空间关系。

动态权值邻接矩阵：通过设定一个初始的邻接矩阵且各元素为$[0,1]$之间的权重值，并在网络训练过程中不断调整邻接矩阵各元素值，公式如下：

$$a_{ij} = a_{ij}^{org} \cdot \delta_{ij} \tag{2-25}$$

其中，a_{ij}^{org} 为动态邻接矩阵元素的初值，$\delta_{ij} \in [0,1]$为可训练的权重值。

自适应邻接矩阵：一些场景下，节点间的相互关系无法预先获得或者并不明显，因此直接使用一组参数通过网络训练的方法获得节点i到节点 j 的权重[32]，邻接矩阵值如下：

$$a_{ij} = SoftMax\left(ReLU\left(\boldsymbol{E}_{1i}\boldsymbol{E}_{2j}^{T}\right)\right) \tag{2-26}$$

其中，$\boldsymbol{E}_{1i}$和$\boldsymbol{E}_{2j}$分别表示节点i和节点j根据语音信号学习到的嵌入特征。

(3) 进化邻接矩阵

在某些情况下，图结构可能会随着时间的推移而演变，比如麦克风阵列中某一个阵元突然停止工作，或麦克风阵列结构发生改变，均会导致图结构发生改变甚至图上一些边缘变得不可用。可以利用一个不断演化的拓扑结构，以捕捉这种动态的空间变化[33][34]。在本文中，我们的邻接矩阵由通过实时训练获得的动态值组成，以表示成对麦克风信号帧之间的实际相移和衰减关系。这种自适应邻接矩阵的定义还可以防止传统语音增强算法由于动态变化和阵列位置测量误差带来的性能下降。

综上，本节基于阵列信号通道间关系包含的目标源信号空间信息、声音传播衰减函数、阵元相互间影响关系，构建了不同类型、不同初值的邻接矩阵，实现在图结构上对图上节点特征进行空间相关性的解析与提取。

2.4 图神经网络搭建

图聚合运算可认为是对多通道信号的幅度谱时频点特征进行加权叠加，本节基于多通道语音信号图聚合运算的图神经网络搭建，利用邻接矩阵对邻居麦克风信号幅度谱进行聚合，提取声阵列信号空间关联性特征，从而更好的对干扰噪声信号去相关，增强目标源信号。

从图谱域的角度按照卷积神经网络对当前像素点求四周相邻像素点加权求和的思路[35]，定义图上第c个麦克风通道第t个时间帧的幅度谱特征为$\boldsymbol{\mathcal{X}}_{c,:,t}$的聚合公式如下：

$$\begin{aligned} \boldsymbol{g}_\theta *_G \boldsymbol{\mathcal{X}}_{c,:,t} &= \boldsymbol{g}_\theta\left(\boldsymbol{L}\right)\boldsymbol{\mathcal{X}}_{c,:,t} \\ &= \boldsymbol{U}\boldsymbol{g}_\theta\left(\boldsymbol{\Lambda}\right)\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\mathcal{X}}_{c,:,t} \end{aligned} \tag{2-27}$$

其中，$*_G$为图卷积操作，$\boldsymbol{g}_\theta$表示图卷积核参数，$\boldsymbol{U} \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$是正交的特征向量矩阵，$\boldsymbol{\Lambda} \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$是对角化特征值矩阵，且$\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} = \boldsymbol{L}$，$\boldsymbol{L}$为拉普拉斯矩阵且$\boldsymbol{L} = \boldsymbol{I}_{\mathrm{C}} - \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{1/2}$，$\boldsymbol{I}_{\mathrm{C}}$是大小为$\mathrm{C}\times\mathrm{C}$的单位矩阵，$\boldsymbol{A}$为声阵列信号邻接矩阵，度矩阵$\boldsymbol{D} \in \mathbb{R}^{\mathrm{C}\times\mathrm{C}}$是一个对角矩阵，表示节点的度即每个节点有多少个邻居节点，即$\boldsymbol{D}_{ii} = \sum_j \boldsymbol{A}_{ij}$。理论上这种对麦克风阵列信号进行图上的卷积是可以保证对其他通道幅度谱特征的聚合的，但是由于$\boldsymbol{U}$的计算复杂度较高，为$\mathrm{O}\left(\mathrm{C}^2\right)$，因此对于由大规模的麦克风阵列组成的图结构，对$\boldsymbol{L}$进行特征值分解的计算资源消耗是难以承受的。

为了避免大规模麦克风阵列场景下 $\boldsymbol{L}$ 的特征值分解导致资源占用过多进而使得算法不可用的情形，采用切比雪夫多项式近似的方法局部化卷积核[36]，公式(2-27)可以表示为：

$$\begin{aligned}\boldsymbol{g}_\theta *_G \boldsymbol{\mathcal{X}}_{c,:,t} &= \boldsymbol{g}_\theta\left(\boldsymbol{L}\right)\boldsymbol{\mathcal{X}}_{c,:,t} \\ &= \sum_{k=0}^{\mathrm{K}-1}\theta_k T_k\left(\tilde{\boldsymbol{L}}\right)\boldsymbol{\mathcal{X}}_{c,:,t}\end{aligned} \tag{2-28}$$

其中，θ 表示切比雪夫多项式系数，$\tilde{\boldsymbol{L}}=\dfrac{2}{\lambda_{\max}}\boldsymbol{L}-\boldsymbol{I}_{\mathrm{C}}$，$\lambda_{\max}$ 表示拉普拉斯矩阵 $\boldsymbol{L}$ 的最大特征根，$T_k\left(x\right)$ 的表达式如下：

$$T_k\left(x\right)=2xT_{k-1}\left(x\right)-T_{k-2}\left(x\right) \tag{2-29}$$

其中，$T_0\left(x\right)=1$，$T_1\left(x\right)=x$。通过递归地计算 $T_k\left(x\right)$，公式(2-27)沿麦克风阵列通道域特征的图卷积可以被局部化为K阶局部卷积，降低了算法的复杂度。

对公式(2-28)进行进一步的简化，约束 K = 2，$T_0\left(\tilde{\boldsymbol{L}}\right)=1$，$T_1\left(\tilde{\boldsymbol{L}}\right)=\tilde{\boldsymbol{L}}$，同时约束 $\lambda_{\max}=2$，则公式(2-28)可以简化表示如下[37]：

$$\boldsymbol{g}_\theta *_G \boldsymbol{\mathcal{X}}_{c,:,t}\approx\theta_0\boldsymbol{\mathcal{X}}_{c,:,t}+\theta_1\left(\boldsymbol{L}-\boldsymbol{I}_{\mathrm{C}}\right)\boldsymbol{\mathcal{X}}_{c,:,t} \tag{2-30}$$

令 $\theta=\theta_0=-\theta_1$ 缓解网络过拟合问题，则公式(2-30)可以进一步简化为：

$$\boldsymbol{g}_\theta *_G \boldsymbol{\mathcal{X}}_{c,:,t}\approx\theta\left(\boldsymbol{I}_{\mathrm{C}}+\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}\right)\boldsymbol{\mathcal{X}}_{c,:,t} \tag{2-31}$$

令 $\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_{\mathrm{C}}$，即对邻接矩阵 $\boldsymbol{A}$ 进行重归一化操作，最终，对第 c 个麦克风通道第 t 个时间帧的幅度谱特征 $\left|\boldsymbol{\mathcal{X}}_{c,:,t}\right|$ 进行图卷积聚合的数学表达如下：

$$\tilde{\boldsymbol{\mathcal{X}}}_{c,:,t}=\boldsymbol{g}_\theta *_G \boldsymbol{\mathcal{X}}_{c,:,t}\approx\theta\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{\mathcal{X}}_{c,:,t} \tag{2-32}$$

其中，$\tilde{\boldsymbol{D}}$ 是 $\tilde{\boldsymbol{A}}$ 的度矩阵。

对麦克风阵列所有通道信号在第 t 个时间帧的幅度谱特征 $\boldsymbol{\mathcal{X}}_{:,:,t}$ 按照公式(2-32)对每个麦克风通道信号进行周围麦克风信号和自身麦克风信号的图聚合操作，并添加激活函数层和 Dropout 层增加网络的非线性能力并失活一部分神经元防止网络过拟合。因此，对声阵列信号幅度谱特征 $\boldsymbol{\mathcal{X}}_{:,:,t}$ 进实现图聚合运算的单层图卷积网络可表达为：

$$\hat{\boldsymbol{\mathcal{X}}}_{:,:,t}=\sigma\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{\mathcal{X}}_{:,:,t}\boldsymbol{W}\right) \tag{2-33}$$

其中，$\boldsymbol{W}\in\mathbb{R}^{\mathrm{C}\times\mathrm{M}}$ 表示具有M个图卷积核的可训练滤波器参数矩阵，单层图卷积网络示意图如下：

考虑图聚合运算的设计原理，不需要构建深层网络，本文发现只需要取图卷积层数约等于阵列中的麦克风个数，即可充分解析各通道信号间的空间关联性。因此，多层图卷积网络可表达如下：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{2-34}$$

其中，$\boldsymbol{H}^{(l)}$ 表示第 l 层图卷积网络的特征输入，$\boldsymbol{H}^{(0)}=\boldsymbol{\mathcal{X}}_{:,:,t}$，$\boldsymbol{W}^{(l)}$ 表示第 l 层图卷积网络的可训练权重矩阵，σ 表示激活函数如 ReLU 等。在图聚合运算理论和邻接矩阵构建的基础上，按照图卷积的思想为每一个节点特征设计可训练的卷积核参数，对邻接矩阵自身特征进行加权并进行重归一化，设计适用于多通道信号处理的图神经网络，可以充分提取阵列信号在空间维度的相关性，并进一步对多通道信号中隐含的噪声分量进行解相关。

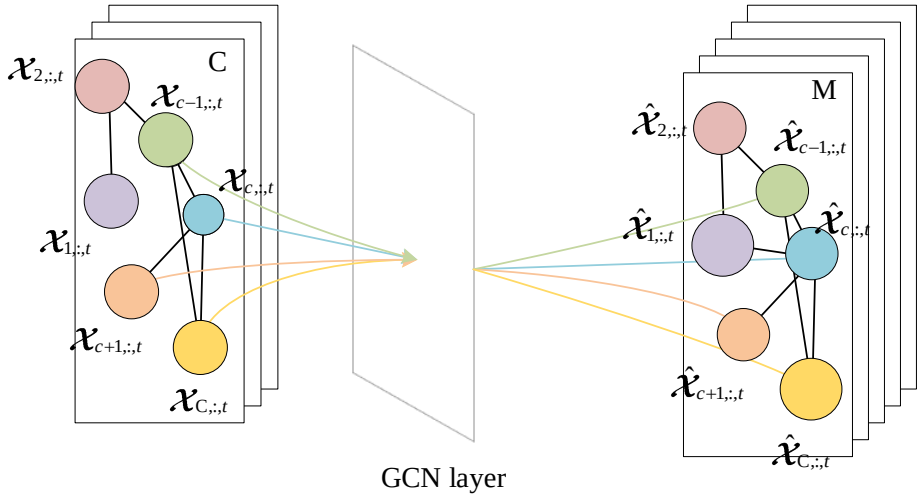


图 2-4 单层图卷积网络示意图

Figure 2-4 Schematic diagram of a single-layer graph convolutional network

2.5 本章小结

本章针对多通道语音增强任务进行了图理论建模与设计。首先根据多通道信号传播模型完成了多通道语音增强场景的建模，分析了基于波束成形的语音增强实质为阵列间信号的加权求和；其次，将阵列信号视为非欧几里得空间中的节点，完成了节点与节点特征的构建并实现了基于阵列信号的图聚合运算；然后，为充分提取阵列信号间的空间相关性特征，完成了不同类型的邻接矩阵构建；最后，搭建了基于多通道信号语音增强问题的图神经网络，完成了多通道语音增强的图理论建模。本节为后续设计时空图卷积语音增强网络，构建空间提取模块、时频提取模块和时空融合模块及针对时空图卷积网络的各项优化奠定了理论基础。

3 时空图卷积语音增强网络构建

在上一章声阵列信号的非欧空间的图理论建模和图聚合运算原理设计的基础上，本章拟融合基于图卷积运算的空间关联性提取方法和基于时频卷积的时域关联性提取方法，构建时空图卷积语音增强网络，提取声阵列信号的空间关联性、时间关联性和时-空依赖性特征，在缺失场景和阵列先验信息的情况下，显著提升目标语音重建质量，抑制噪声和干扰。具体包括：首先，对声阵列信号的空间关联性、时间关联性和时-空依赖性进行分析；其次，设计构建时空图卷积语音增强网络，主要包括空间关联性提取模块、时空信息融合模组和多通道融合模块等；最后，针对所设计的时空图卷积语音增强网络，进行了多种对比算法、多种评估方法、多种邻接矩阵和多种数据集的对比实验，从多个维度证明了本章提出的时空图卷积语音增强网络算法在多场景中的语音增强有效性和鲁棒性。

3.1 声阵列信号时空关联性分析

3.1.1 空间关联性分析

不同于单麦克风采集声音信号时无法获得信号的空间信息，麦克风阵列在采集声音信号时可以根据不同麦克风采集到信号的差异获得声源信号的空间位置信息。空间中点声源在某一时刻的信号在传播过程中经过不同的时延和衰减被麦克风阵列中不同的阵元接收，可以根据不同阵元在某时刻采集到的信号，结合阵列形状、阵列间距信息获取到点声源相对于阵元中心的空间方位；在多源嘈杂环境中，麦克风阵列在工作时往往同时接收多个点声源信号，不同点声源自身具有不同的能量且相对于麦克风阵列具有不同的空间位置，此时麦克风阵列接收到的信号不仅包含了声源信号本身的空间信息，也包含了不同声源的相对位置信息；当考虑到声源移动的情况，麦克风阵列在t_0时刻和在t_1时刻接收到的阵列信号具有不同的空间关联性；室内声场环境中，声源信号在传播过程中受到房间墙面、房间中物体等的反射，麦克风阵列最终接收到的是包含不同传播路径的含混响信号。

因此，结合以上分析可知，声阵列信号的空间关联性指的是不同麦克风采集到的信号在同一时刻彼此高度相关，且具有局部优先(即临近麦克风的信号有强相似性)、多点关联(即各通道信号均可依据阵列流形特征由其他通道信号聚合而来)和全局结构恒定(即阵列拓扑长时不变)的特点。

临近麦克风的信号有强相似性，且不同的空间声源位置会导致阵列中两个阵元具有不同的空间相关性。下面以均匀线阵为例，，对这种关联性进行量化表达。假设阵列中第一个麦克风为参考麦克风，声波到达阵列的入射角为θ，则：

$$\bar{X}\left(f;\boldsymbol{r}_s,\boldsymbol{r}_c\right)=\frac{X\left(f;\boldsymbol{r}_s,\boldsymbol{r}_c\right)}{X\left(f;\boldsymbol{r}_s,\boldsymbol{r}_1\right)} \tag{3-1}$$

因此均匀线性阵列的响应向量$u\left(\boldsymbol{r}_s\right)$可以表示为：

$$\begin{aligned} u\left(\boldsymbol{r}_s\right)&=\left[\bar{X}\left(f;\boldsymbol{r}_s,\boldsymbol{r}_1\right),...,\bar{X}\left(f;\boldsymbol{r}_s,\boldsymbol{r}_{\mathrm{C}}\right)\right]^{\mathrm{T}} \\ &=\left[1,...,\exp\left(-j2\pi\frac{\left|\boldsymbol{r}_s-\boldsymbol{r}_{\mathrm{C}}\right|-\left|\boldsymbol{r}_s-\boldsymbol{r}_1\right|}{\mu}\right)\right]^{\mathrm{T}} \end{aligned} \tag{3-2}$$

则阵列中第i个阵元和第 j 个阵元的空间关联性[38]可表示为：

$$\begin{aligned} \rho\left(i,j\right)&=\mathrm{E}\left[u_i\left(\boldsymbol{r}_s\right)u_j\left(\boldsymbol{r}_s\right)^{\mathrm{H}}\right] \\ &=\int_{\boldsymbol{r}_s\in\mathrm{FOV}}u_i\left(\boldsymbol{r}_s\right)u_j\left(\boldsymbol{r}_s\right)^{\mathrm{H}}\,p\left(\boldsymbol{r}_s\right)d\boldsymbol{r}_s \end{aligned} \tag{3-3}$$

其中，$p\left(\boldsymbol{r}_s\right)$是声源信号沿入射方向的角能量概率密度函数，$p\left(\boldsymbol{r}_s\right)$的计算方法及$\rho\left(i,j\right)$的计算结果间附录 B。由公式(3-3)可知，不同的空间声源位置会导致阵列中两个阵元具有不同的空间相关性，从而使得接收到的阵列信号不同通道间具有不同的空间关联。

综上，多通道信号中含有丰富的多维空间关联性特征，可以表征声源位置、阵列结构、声源速度、声学环境等多种信息。在陌生的复杂应用环境中，如何利用这些空间特征信息来进行多通道语音增强，提取目标分量，是提升多通道语音增强算法性能上限的关键。本小节从三维空间声源的角度分析了多种场景下阵列接收信号时存在的空间关联性，为后续 3.2.2 节构建空间关联性提取模块提供理论依据。

3.1.2 时间关联性分析

声阵列信号的时域关联性具有多尺度相关(即混杂长时平稳背景声、同类型短时跳发噪声、同一说话人的短时语音发声音素特征等)和时变性(即信号帧之间的关联关系会随时间发生变化)的特点。本节从语音信号本身特性及多通道信号接收特性的角度分析声阵列信号的时间关联性。

语音信号随时间变化表现出短时平稳特性，而在多通道信号中，不同通道信号相互之间又有与时间相关的不同时刻通道信号的相似性。语音由声带振动以及空气湍流作用于声道产生，音素是语音的最小单位，人在说话时每个字/词由一个/多个音素组成，每个音素都会持续一段时间(50-200 ms 之间)。语音音素存在概率

示意图如图 3-1 所示，横轴表示时间(s)，纵轴表示音素类型，图中白色深浅表示音素存在概率的强弱，白色长度表示音素持续时间，因此对由音素组成的语音而言，其具有短时平稳特性，通常依据此特性对语音信号进行分帧处理，即不同时间帧的语音信号存在着时间上的关联性，表现为一个语音帧信号的频谱特性是稳定的，前后相邻的语音帧信号的频谱变化是缓慢的。

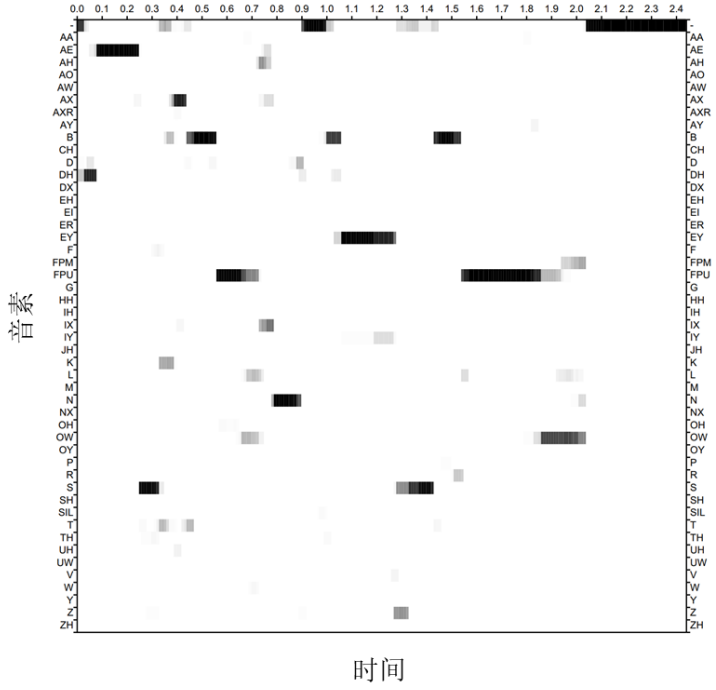


图 3-1 一段语音得音素存在概率图[39]

Figure 3-1 A schematic diagram of probability of phoneme existence in a speech[39]

此外，考虑到空间中的声源信号经过不同的幅度衰减和时间延迟被阵列中不同的麦克风采集，以下给出不同麦克风通道信号时间相关性的定量分析，则第i个通道t_1时间帧和第 j 通道t_2时间帧信号的时间互相关性系数定义为：

$$\rho_{ij}\left(t_1,t_2\right)=\frac{\sum_{t=0}^{\mathrm{T}}\boldsymbol{\mathcal{X}}_{i,f,t_1}\boldsymbol{\mathcal{X}}_{j,f,t_2}}{\sqrt{\sum_{t=0}^{\mathrm{T}}\left(\boldsymbol{\mathcal{X}}_{i,f,t_1}\right)^2\sum_{t=0}^{\mathrm{T}}\left(\boldsymbol{\mathcal{X}}_{j,f,t_2}\right)^2}} \tag{3-4}$$

体现出不同时间两通道信号具有的不同时间相关性。

本节分析了语音信号本身的时间关联性，并从声阵列信号的角度分析了不同通道信号之间的时间关联性，揭示了声阵列信号内在的时间依赖关系，为下一步构建时频信息提取模块奠定理论基础。

3.1.3 时-空依赖性分析

语音信号的空间关联性和时间关联性并不是独立的，其相互影响和依赖。从信号的角度看，源信号沿直接路径与沿反射的其他路径被麦克风接收，存在着接收时间上的差异，而沿不同路径被麦克风接收又表现出空间上的差异，表现为点声源入射方向与反射声源入射方向的不同；从图节点特征上看，在一个时间帧上每个节点特征都可以直接影响到其邻居节点特征，由于时间序列的时间相关性，当前时间帧每个节点特征也可以在下一个时间帧直接影响自身特征，此外，由于时空信息是同步的，当前时间帧每个节点特征可以在下一个时间帧直接影响其邻居节点特征[40]，如图 3-2 所示。因此需要针对含噪声阵列信号中隐含的不同尺度、不同维度的信号关联关系进行深度刨析。

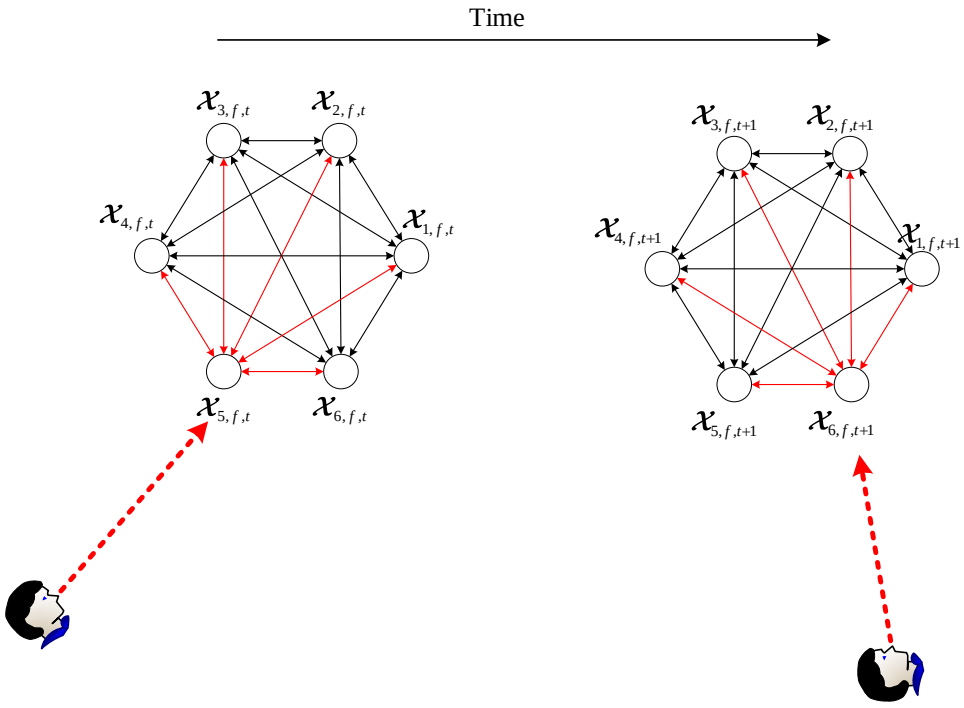


图 3-2 声阵列信号的时-空依赖性

Figure 3-2 Spatial-temporal dependence of acoustic array signals

单独对空间关联性或者时频关联性建模无法有效利用声阵列信号的空间特征和时间特征之间的潜在相互作用，这会降低多通道语音增强系统得性能。因此，在后续章节中，本文将空间关联性提取模块和时频关联性提取模块相融合，组合形成时空信息融合模组来联合捕获声阵列信号时空依赖性，更好的对多源信号进行解相关。

3.2 时空图卷积语音增强网络

本节提出时空图卷积语音增强网络，来全面解析声阵列信号的空间关联性、

时频关联性和时-空依赖性特征，从而实现高效且鲁棒的目标语音增强。本节首先给出网络总体框架设计；其次详细介绍了基于图卷积运算的空间关联性提取模块和基于标准二维卷积的时频关联性提取模块；接着介绍了提取声阵列信号时-空依赖关系的时空信息融合模组；最后多通道融合模块将输出的多通道信号融合为单通道增强信号。

3.2.1 网络总体框架设计

以麦克风阵列采集到的含噪语音信号幅度谱特征 $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{C\times F\times T}$ 为输入，多通道语音增强问题可以表示为网络学习一个映射函数 $\hbar(\bullet)$: $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{C\times F\times T} \xrightarrow{\hbar(\bullet)} \hat{\boldsymbol{Y}} \in \mathbb{R}^{F\times T}$，其中 $\hat{\boldsymbol{Y}}$ 是干净语音幅度谱 $\boldsymbol{Y}$ 的估计值。

以麦克风阵列采集到的含噪语音信号幅度谱特征 $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{C\times F\times T}$ 为输入，多通道语音增强问题可以表示为网络学习一个映射函数 $\hbar(\bullet)$: $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{C\times F\times T} \xrightarrow{\hbar(\bullet)} \hat{\boldsymbol{Y}} \in \mathbb{R}^{F\times T}$，其中 $\hat{\boldsymbol{Y}}$ 是干净语音幅度谱 $\boldsymbol{Y}$ 的估计值。如图 3-3 所示，本节基于级联的 ST 模组和通道融合模块构建的时空图卷积语音增强网络，建立多通道含噪语音信号幅度谱特征到估计的干净语音信号幅度谱特征的映射函数，通过空间 GCN 模块提取声阵列信号空间关联性，通过时频 CNN 模块提取各通道信号时频关联性，实现无需任何阵列结构、目标声源定位和声学场景等先验信息的声阵列信号语音增强，显著提升网络输出语音的 SNR 和 SIR。

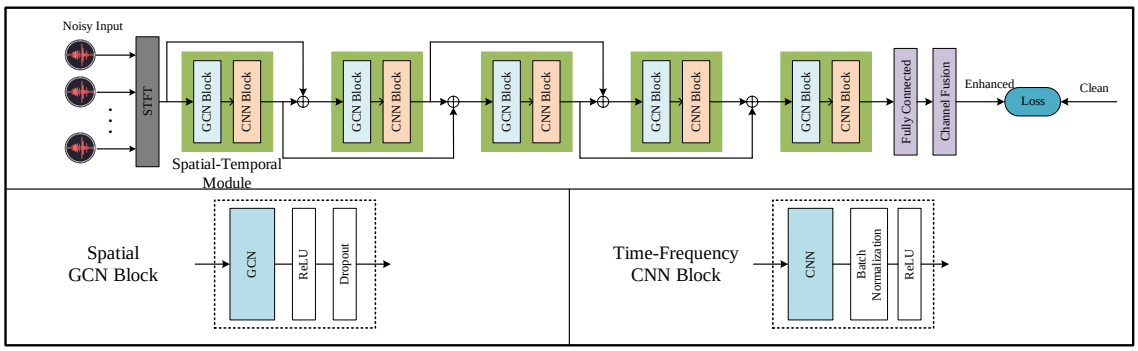


图 3-3 语音增强网络整体结构

Figure 3-3 Overall structure of speech enhancement network

本文提出的时空图卷积语音增强网络框架主要由多个级联的时-空(ST)依赖性融合模组和一个通道融合模块组成，其中 ST 模组用于提取多通道含噪语音信号的空间和时频信息，ST 模组中的空间 GCN 模块提取与不同麦克风位置相关的多通道信号之间的空间依赖性，时频 CNN 模块用于捕获每个麦克风通道信号的视频特征，进一步将语音与噪声去相关，模组之间添加跳连接增加网络的泛化性能。多

通道含噪语音信号幅度谱特征由多个级联的ST模组进行空间关联性、时频关联性和时空依赖性特征提取，通道融合模块以最后一个ST模组的结果为输入，对没个麦克风通道时频点特征添加一个自适应可训练权重，时频点特征沿通道域进行加权叠加得到估计的目标源语音信号幅度谱 $\hat{\boldsymbol{Y}}$ 。网络损失函数设置为最小均方误差(Mean Squared Error, MSE)损失函数，即计算 $\hat{\boldsymbol{Y}}$ 与 $\boldsymbol{Y}$ 的均方误差，网络根据损失函数值进行反向传播，对每一层网络参数进行梯度下降更新网络参数值。

3.2.2 基于图卷积运算的空间关联性提取模块

对声阵列信号空间关联性的特点分析决定了本节拟采用非欧空间的图卷积聚合运算来进行阵列结构化特征的提取。结合 2.4 中构建的多通道信号语音增强图神经网络结构，构建基于图聚合运算的时空图卷积模块，来进行含噪阵列信号幅度谱的空间关联性特征提取。

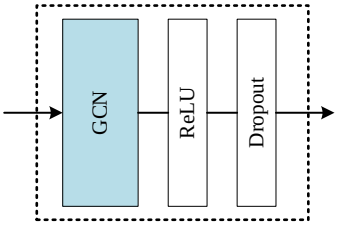


图 3-4 空间 GCN 模块结构图

Figure 3-4 Diagram of Spatial GCN block structure

传统基于信号模型的方法在实际应用中往往由于声场环境的复杂多样性，传统基于信号模型的方法对声源传播过程中的多径反射及混响估计不足，与理论传播模型的假设出现偏差。同时由于麦克风阵列测量误差及声源定位误差的影响，实际应用过程中未能充分利用信号的空间关联性。将麦克风阵列构建为图结构，麦克风通道信号的图聚合运算在处理阵列中的一个通道信号时对其他有关联的麦克风通道信号进行聚合，即可以根据接收到的多通道信号的幅度和相位差来估计与分布的麦克风位置相关的目标语音分量的空间依赖性。在 2.4 节中，我们根据 t 时刻的图信号幅度谱特征 $\boldsymbol{\mathcal{X}}_{:,:,t}$ 搭建了基于阵列信号处理的图神经网络 $g_\theta *_{\mathrm{G}} \{\boldsymbol{\mathcal{X}}_{:,:,t}\} = \tilde{\boldsymbol{D}}^{-1/2}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}^{1/2}\boldsymbol{\mathcal{X}}_{:,:,t}\boldsymbol{W}$ ，为了使图卷积网络为邻居节点信息的非线性聚合函数分配更大的灵活性，在图卷积层后面添加了 ReLU 激活函数层和 Dropout 层，最终构建了空间图卷积模块，同时，网络模型的数据输入为包含了 T 个时间帧的多通道含噪信号幅度谱为：

$$\boldsymbol{\mathcal{X}} = \left[\boldsymbol{\mathcal{X}}_{:,:,1},...,\boldsymbol{\mathcal{X}}_{:,:,\mathrm{T}}\right] \tag{3-5}$$

则空间图卷积模块数学表达如下：

$$\boldsymbol{\Gamma}\{\boldsymbol{\mathcal{X}}\}=Concat_{\mathrm{T}}\left(Dropout\left(ReLU\left(g_{\theta}*_{\mathrm{G}}\{\boldsymbol{\mathcal{X}}_{:,:,t}\}\right)\right)\right) \tag{3-6}$$

其中，$\boldsymbol{\Gamma}\in\mathbb{R}^{\mathrm{C}\times\mathrm{M}\times\mathrm{T}}$ 表示空间图卷积模块的输出，$Concat_{\mathrm{T}}(\bullet)$ 网络在每一时刻的计算结果沿时间帧 $t=1,...,\mathrm{T}$ 进行拼接操作。空间图卷积模块网络结构示意图如图 3-4 所示。

因此，基于图卷积运算的空间关联性提取模块实现了对一个时间帧图信号的图聚合运算，并添加激活函数层和池化层提升网络模块的非线性实现对图信号进行空间关联性进行提取，将一段语音信号的多个时间帧幅度谱信号沿时间维度进行拼接有效提取阵列信号幅度谱特征在一段时间内的通道间空间关联性。

3.2.3 时空信息融合模组

除了接收到的多通道信号与分布式麦克风位置相关的空间依赖性外，每个通道的信号时频域中还存在多种相关性，这些相关性来源于声学场景中的静态背景噪声、突发干扰和说话者的音素特征。此外，多通道噪声信号的空间依赖性和时频相关特征相互影响。因此，我们设计了级联的 ST 模组来提取接收到的多通道噪声信号的时变时空特征，用于语音分量的盲增强。

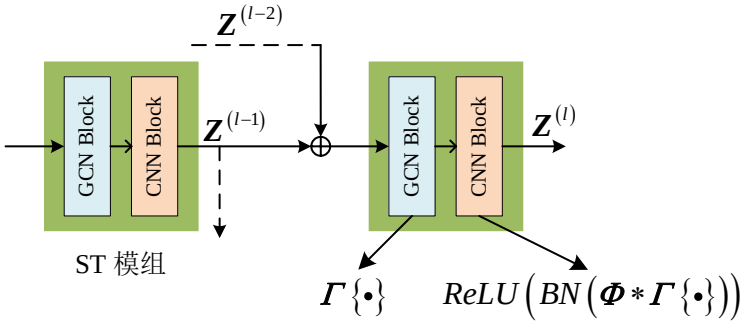


图 3-5 声阵列信号在级联的 ST 模组中传播关系

Figure 3-5 Propagation relationship of acoustic array signals in cascaded ST modules

考虑到声阵列信号的空间关联性具有长时不变的特点，本文提出的 ST 模组由一个前端空间 GCN 模块(在第 3.3 节中讨论)和一个时频 CNN 模块组成，用于提取每个麦克风通道中信号的时频相关性。基于卷积运算的时频 CNN 模块，对于每一个麦克风通道的幅度谱特征按照标准 CNN 的方法进行卷积，同时添加 Batch Normalization 层和 ReLU 层以加速网络训练并增加网络的非线性能力，从而更有效地提取阵列信号的时频相关性。ST 模组的数学表达式为：

$$\boldsymbol{Z}=ReLU\left(BN\left(\boldsymbol{\Phi}*\boldsymbol{\Gamma}\{\boldsymbol{\mathcal{X}}\}\right)\right) \tag{3-7}$$

其中，$*$ 表示标准二维卷积，$\boldsymbol{\Phi}$ 为时频域的二维卷积核参数。

如图 3-5 所示，多个 ST 模组级联以增加网络深度，提升了网络对声阵列信号时空频特征提取能力，同时添加跳连接，第 l 个 ST 模组的输出 $\boldsymbol{Z}^{(l)}$ 可表示如下：

$$\boldsymbol{Z}^{(l)} = ReLU\left(BN\left(\boldsymbol{\Phi} * \boldsymbol{\Gamma}\left\{\boldsymbol{Z}^{(l-1)} \oplus \boldsymbol{Z}^{(l-2)}\right\}\right)\right) \tag{3-8}$$

其中，$\boldsymbol{Z}^{(0)} = \boldsymbol{\mathcal{X}}$，$\boldsymbol{Z}^{(1)} = ReLU\left(BN\left(\boldsymbol{\Phi} * \boldsymbol{\Gamma}\left\{\boldsymbol{Z}^{(0)}\right\}\right)\right)$，$l \in \{2,...,\mathrm{L}\}$，L 表示本文提出的 STGCSEN 的级联的 ST 模组总数。通过实验发现，L 的最优值为C-1，这与聚合邻居节点的数量一致，可以很好的用图卷积理论解释。

综上，本节通过构建时空信息融合模组，对声阵列信号时空依赖关系进行建模，从通道域、时间帧和频点维度提取声阵列信号空间关联性与时频关联性的相互依赖关系。阵列信号首先经过空间 GCN 模块获取到空间关联性，再将空间信息提取模块处理后的信号经时频 CNN 模块处理，融合提取阵列信号的时空依赖性。多个 ST 模组级联增加了网络深度，提升了网络的时空依赖性特征表达能力，级联结构增加了网络的泛化能力。

3.2.4 多通道融合模块

针对多个级联的 ST 模组输出的多通道信号时、空、频多维特征，需要进行通道间信号的融合以得到估计的单通道目标源信号。本节基于多通道信号时频点特征自适应权重叠加的思想设计了通道融合模块。

如图 3-6 所示，L 个级联的 ST 模组有效提取了阵列信号的空间关联性和时频关联性，之后一个全连接层和一个通道融合模块将处理后的多通道信号按多维关联性特征进行目标语音估计，数学表示如下：

$$\hat{\boldsymbol{Y}} = \sum_{c=1}^{\mathrm{C}} \boldsymbol{\varpi}^{c} \odot \left(\mathcal{F}\left\{\boldsymbol{Z}_{c,:,:}^{(\mathrm{L})}\right\}\right) \tag{3-9}$$

其中，$\hat{\boldsymbol{Y}} \in \mathbb{R}^{\mathrm{F}\times\mathrm{T}}$ 估计的目标源语音幅度谱，$\odot$ 表示哈达玛乘积，$\mathcal{F}\{\bullet\}$ 表示全连接层，$\{\boldsymbol{\varpi}^{1},...,\boldsymbol{\varpi}^{\mathrm{C}}\}$ 是每个麦克风通道恢复语音频谱的特征权重矩阵。

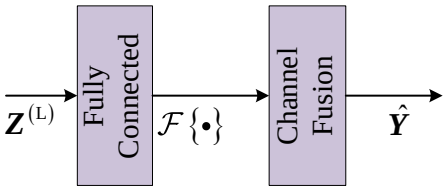


图 3-6 通道融合模块结构图

Figure 3-6 Diagram of channel fusion module structure

本节针对多个级联得 ST 模组最终输出的多通道信号时、空、频特征，设计了通道融合模块，为每个麦克风通道的每个时频点分配自适应权重，多通道时频点

特征沿通道域加权融合获得估计的单通道目标源语音信号幅度谱特征，完成整个时空图卷积网络语音增强过程。

3.3 实验设计及结果分析

3.3.1 数据集与评价指标

(1) 数据集

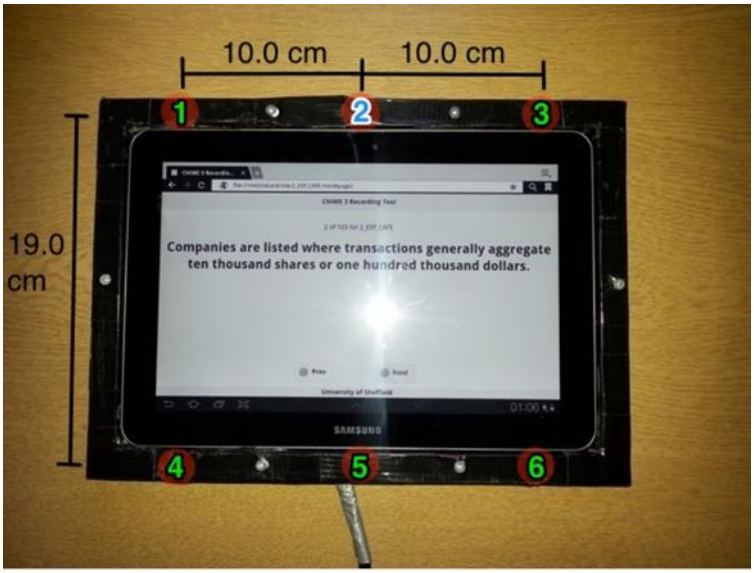


图 3-7 麦克风阵列结构图[41]

Figure 3-7 Microphone array structure diagram[41]

本文实验使用公开数据集 CHiME3[41]。CHiME3 数据集包含大约 342 小时的英语语音和约 50 小时的嘈杂环境噪声，其应用场景侧重于现实环境中的远场麦克风阵列语音识别/增强。干净语音基于华尔街日报(WSJ0)语料库，语料中说话人相互独立，噪声数据来自于记录的四种真实嘈杂环境，包括公共汽车(BUS)、咖啡馆(CAF)、步行区(PED)和街道路口(STR)，含噪语音数据通过人为地将干净语音数据和噪杂的背景噪声数据按照如图所示的录音设备提供的麦克风阵列结构进行合成。训练集包括 7138 段由共 83 位说话人在四种嘈杂环境中合成的含噪语音，验证集包括 1640 段训练集之外的 4 位说话人在四种嘈杂环境中合成的含噪语音，测试集包括 1320 段其他的 4 位说话人在四种嘈杂环境中合成的含噪语音。

为了充分利用现有数据，我们对原数据集中的数据进行了裁剪，将不同长度的音频均裁剪为长度为 3 s 的音频，并将含噪语音与原干净语音进行配对，最终生成了 11716 个含噪-干净语音对作为训练集，2930 个含噪-干净语音对作为测试集，

BUS、CAF、PED、STR 四种场景的测试集数量分别为 752、745、684、749。数据集中所有音频信号采样率均为 16000 Hz。

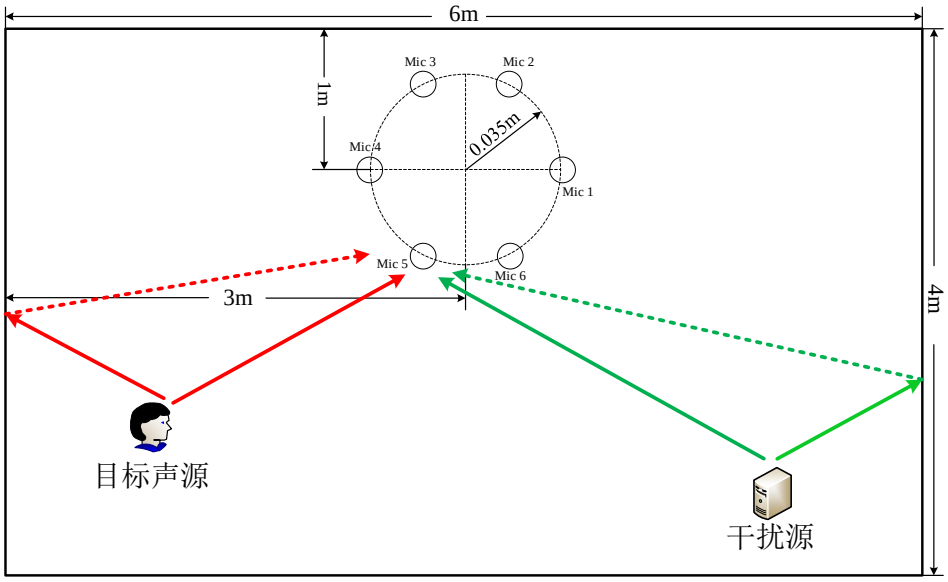


图 3-8 室内环境及麦克风阵列设置俯视图

Figure 3-8 Top view of indoor environment and microphone array setup

实验中使用的第二个数据集 DNS-Set 来自 ICASSP 2022 Deep Noise Suppression (DNS) Challenge 4 提供的开源数据集[42]，其干净语音数据来自 LibriVox[43]，一个公共有声读物数据集，包含 2150 个说话人且说话人男女比例大致均等，干净语音总时长超过 500 小时；噪声语音数据来自 AudioSet[44]和 Freesound[45]，AudioSet 是从 2084320 个 YouTube 视频中提取出的共 527 个标签的噪声片段，每个噪声片段为 10 秒，Freesound 包含约 10000 个噪声片段。随机选取语音数据和噪声数据并将其截断为长度均为 3 秒的音频，同时进行活动性检测、截幅等预处理操作，排除掉含有大量静音段的音频和部分语音有效信息丢失的音频(当音频时域幅值大于 0.99 时认为语音在录制时出现了截幅，丢失了部分有效信息)。最终获得了长度为 3 秒的干净语音数据和噪声数据各 20580 段，总计音频超过 34 小时。按照图像源镜像(Image Source Method, ISM)[46]的方法计算阵列中麦克风的室内冲激响应(Room Impulse Response, RIR)，房间及麦克风阵列的设置如图 3-8 所示，设置房间的长、宽、高分别为 6、4、3 米，麦克风阵列为均匀 6 麦克风圆阵，半径 0.035 米，阵列中心距离 6 米墙面一侧 1 米，距离 4m 墙面 3 米，高度始终为 1 米，所有麦克风均为全向麦克风。说话人声源与点噪声源在房间中随机位置出现，且源位置均距离麦克风阵列中心距离大于 0.3 米，距离各墙面距离大于 0.3 米。使用 Python 库 Pyroomacoustics[47]按照 ISM 方法实现室内空间声学模拟，混响时间 T_{60} (能量下降至 60dB 所需时间)为随机值，范围在 $[0.2, 0.8]$ 秒之间，根据混响时间、房间尺寸和声速计算所需的能量吸收系数和最大声源阶数。每一次随机从干净语

音和点源噪声中各选取一条音频，按照相同的 RIR 分别卷积干净语音和点源噪声生成多通道空间音频，将空间干净语音和空间点源噪声按不同的信噪比叠加生成最终的含噪语音，生成含噪语音过程中 SNR 范围是$[-5,10]$dB。

(2) 评价指标

本文采用语音质量感知评估(Perceptual Evaluation of Speech Quality, PESQ)[48]和短时语音可懂度(Short-Time Objective Intelligibility, STOI)[49]两种客观语音质量评估方法和带有隐藏参考和锚点的多激励测试(Multi-Stimulus Test with Hidden Reference and Anchor, MUSHRA)[50]主观语音质量评估方法对我们的模型进行评估。

PESQ 是语音质量评估中最常用的客观评价指标，用于衡量重建目标语音的失真度，其测试结果与主观听觉测试有较高的相关性，使用时提供重建的目标语音及对应的干净语音，取值范围是$[-0.5,4.5]$，分值越高表示语音质量越好。STOI 评价指标用于评价语音的可懂度，使用时提供重建的目标语音及对应的干净语音，取值范围是$[0,1]$，分值越高表示语音可懂度越好。MUSHRA 是一种主观评价语音质量的测试方法，分值范围$[0,100]$，分值越高表示语音的主观感知质量越好。在我们的实验中，8 名未接受任何专业声学听力训练的人对不同模型不同场景下重建的目标语音给出他们最直观的主观评价分数，在每种场景下，我们为每个模型选取 5 个音频样本，因此每名试听者需要对总计 80 个音频样本进行主观听力评分，所有试听者在试听之前，未获得任何关于音频样本和模型的先验信息，音频样本以 Sample1、Sample2、Sample3、Sample4 和 Sample5 表示，模型以 Model1、Model2、Model3、Model4 表示。

3.3.2 基线模型与实验设置

为了评价我们设计的 STGCSEN 模型在不同场景下语音增强的有效性，本文选择了三种目前在多通道语音增强领域取得较好效果的模型作为对比基线模型。

(1) CNN[51]，由全卷积网络组成的编码器-解码器型语音增强模型，5 个卷积块作为编码器，5 个反卷积块作为解码器。我们修改了模型的输入，使模型原来的的单通道输入现在可以处理多通道信号，模型的输入为含噪语音的幅度谱，输出为估计的干净语音幅度谱，时频域的幅度谱上卷积核大小设置为$(3,2)$，步长设置为$(2,1)$，编码器中每一个卷积块的输出信道为$\{16,32,64,128,256\}$，解码器与编码器对称，每一个反卷积块的输入信道为$\{256,256,128,64,32\}$，编码器中的每个卷积块与对应的解码器中的每个反卷积块之间添加跳连接。

(2) LSTM-IPD[52]，由三个 LSTM 层和一个全连接层组成的模型，每个 LSTM 层包含 512 个隐藏层单元，全连接层包含 514 个神经元，模型的输入是含噪语音

频谱的实部、虚部和通道间相位差，输出为估计的目标语音的实部和虚部。

(3) UNet-GCN[29]，由卷积块和图卷积层组成的 UNet 型多通道语音增强模型，6个卷积块和6个反卷积块分别是编码器和解码器结构，编码器和解码器之间嵌入了两层图卷积层，模型的输入、输出、卷积核大小和步长与(1)中 CNN 结构相同，图卷积层的图结构为 6 节点图，表示 6 个麦克风通道，节点间关系由大小为6×6的全 1 邻接矩阵表征。

本文的 STGCSEN 模型采用 5 个级联的 ST 模组，每个模组之间添加跳连接，卷积核大小为$(5,2)$，步长设置为$(2,1)$，其中第三个 ST 模组的卷积核步长分别为$(3,3)$和$(1,1)$，最后一个 ST 模组中的时频相关性提取模块使用 ELU 激活函数。

对输入的 6 通道含噪语音做 STFT 变换将时域信号变换为频域信号时，STFT 窗长为512个采样点(32 ms)，并对512个采样点做FFT变换，每次滑动距离为256个采样点，窗函数选择为汉明窗。所有模型均使用Pytorch实现，用于训练的优化器为 Adam，学习率设置为固定值 0.006，其动量参数设置为 0.9 和 0.999，Batch Size 设置为 12。

3.3.3 实验结果分析

本节设计了四组实验，分别是多算法对比实验、主观语音评价实验、多邻接矩阵对比实验和多数据集对比实验。表中数据粗体表示对比实验结果中最优的结果，数据加下划线表示对比实验结果中取得了效果为次优(第二)或者较优(第三)的结果。

3.3.3.1 多算法对比实验

四种不同算法模型在四种不同真实场景(BUS, CAF, PED, STR)下的 PESQ 分值和 STOI 分值对比分别如表 3-1 和表 3-2 所示。本文提出的 STGCSEN 模型其 PESQ 分值在所有嘈杂场景中都优于基线模型，在各种场景中综合的 STOI 分值最优。具体来讲，本文提出的模型在 PESQ 评估上相比于基线模型取得了平均约 11%-19%的性能提升，在 STOI 指标评估上，我们的方法在 CAF 和 STR 两种场景中取得了最优的表现，同时，四种嘈杂噪声场景下我们的综合 STOI 指标达到了最优，说明我们的方法在不同场景中均有稳定的表现。综上，通过有效提取多通道信号的空间和时间关联性，本文所提出的 STGCSEN 模型在语音增强方面实现了相当或最优的性能，并对不同的噪声场景表现出更强的适应性。

表 3-1 不同模型在四种噪声环境下的 PESQ 对比

Table 3-1 PESQ comparison of different models in four noise environments

Model	BUS	CAF	PED	STR	Avg.
Noisy (channel 1)	1.28	1.15	1.18	1.26	1.22
CNN	1.78±0.26	1.54±0.21	1.62±0.24	1.75±0.30	1.67±0.27
LSTM-IPD	1.91±0.06	1.49±0.20	1.51±0.26	1.73±0.34	1.66±0.35
UNet-GCN	1.66±0.20	1.44±0.17	1.49±0.19	1.64±0.26	1.56±0.18
STGCSEN	**2.00±0.29**	**1.75±0.25**	**1.74±0.27**	**1.74±0.29**	**1.86±0.29**

表 3-2 不同模型在四种噪声环境下的 STOI 对比

Table 3-2 STOI comparison of different models in four noise environments

Model	BUS	CAF	PED	STR	Avg.
Noisy (channel 1)	0.91	0.81	0.81	0.88	0.85
CNN	0.92±0.03	0.89±0.04	**0.89±0.04**	0.91±0.03	**0.90±0.04**
LSTM-IPD	**0.93±0.06**	0.85±0.07	0.83±0.10	0.89±0.06	0.87±0.08
UNet-GCN	0.92±0.03	0.88±0.05	0.87±0.05	0.91±0.03	0.90±0.06
STGCSEN	0.92±0.06	**0.89±0.06**	0.88±0.06	**0.92±0.06**	**0.90±0.06**

3.3.3.2 主观语音评价实验

基于 MUSHRA 的主观实验，各算法在不同场景下的主观听力评估分值如表 3-3 所示。本文所提出的模型在各种场景下均有最优的表现，同时主观听力综合评估指标也明显优于其他三种基线模型，在四种真实场景和综合评估中，我们的模型有更低的均方根误差，说明模型在各种语音场景下对语音质量的提升都非常稳定，显示出模型在基于人类听觉机理上更好的语音质量和可懂度，模型的鲁棒性更好。同时根据试听者在听觉感知测试后的反馈看，我们的模型与干净的语音有更多的相似性，在无声段中听起来更加自然。

不同算法输出的重建目标语音的语谱图对比如图 3-9 所示，CNN、LSTM-IPD 和 UNet-GCN 三种算法对于平稳噪声都有一定的抑制效果，其中 LSTM-IPD 的抑制平稳噪声时语音增强效果最好，但三种算法对于非平稳噪声的抑制效果均较差，LSTM-IPD 对非平稳噪声几乎没有进行抑制。本文提出的 STGCSEN 在抑制平稳噪声和非平稳噪声上均有比较明显的效果，增强后的语音更加接近干净语音。

表 3-3 不同模型在四种噪声环境下的主观听力测试结果对比

Table 3-3 Comparison of subjective hearing test results of different models in four noise environments

Model	BUS	CAF	PED	STR	Avg.
CNN	78.85±6.01	79.73±4.73	81.63±4.79	81.85±5.56	80.52±5.27
LSTM-IPD	83.85±7.59	79.25±7.68	74.03±6.66	83.98±6.47	80.28±7.10
UNet-GCN	77.50±4.64	76.58±5.50	79.20±6.74	81.75±5.87	78.76±5.69
STGCSEN (ours)	**84.30±4.20**	**85.05±4.19**	**83.35±6.43**	**86.75±5.90**	**84.86±5.18**

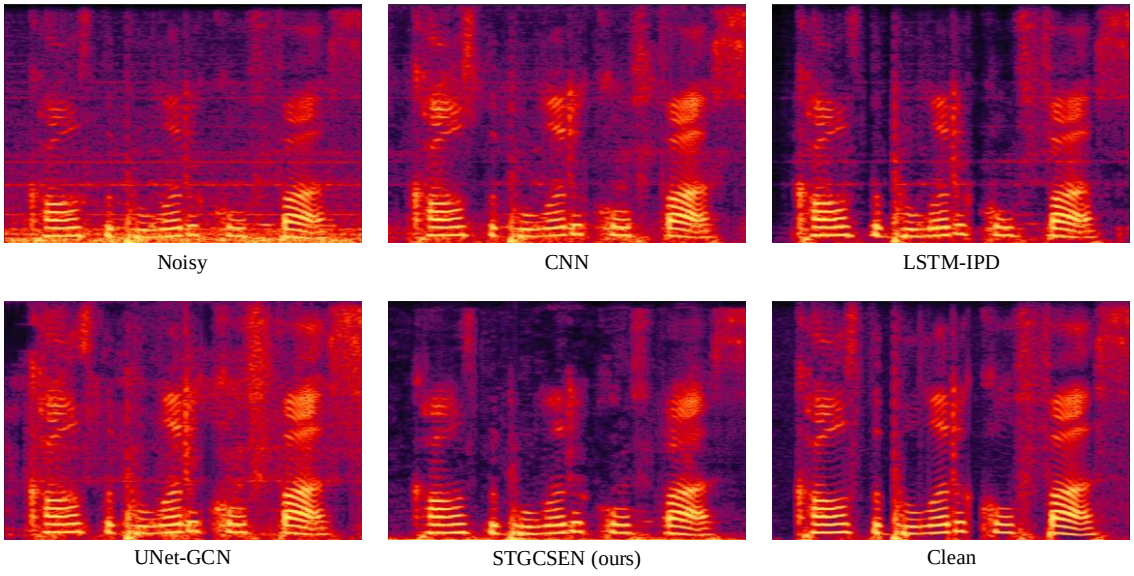


图 3-9 不同模型增强语音语谱图对比

Figure 3-9 Comparison of spectrograms of enhanced speech with different models

3.3.3.3 多邻接矩阵对比实验

本节根据麦克风阵列不同麦克风间距设计了距离衰减邻接矩阵、指数距离衰减邻接矩阵和反对称邻接矩阵。在本组实验中，STGCSEN 表示全 1 固定邻接矩阵的时空图卷积语音增强网络，STGCSEN-Fix*表示使用不同的固定邻接矩阵，STGCSEN-Dym*表示使用动态邻接矩阵，且以不同的固定邻接矩阵作初值。根据 2.3 节，STGCSEN-FixA1 中的邻接矩阵值为距离衰减邻接矩阵值，具体的邻接矩阵设置如下：

$$\boldsymbol{A}_{\text{FixA1}}=\begin{bmatrix} 0 & \frac{1}{10} & \frac{1}{20} & \frac{1}{19} & \frac{1}{\sqrt{10^2+19^2}} & \frac{1}{\sqrt{19^2+20^2}} \\ \frac{1}{10} & 0 & \frac{1}{10} & \frac{1}{\sqrt{10^2+19^2}} & \frac{1}{19} & \frac{1}{\sqrt{10^2+19^2}} \\ \frac{1}{20} & \frac{1}{10} & 0 & \frac{1}{\sqrt{19^2+20^2}} & \frac{1}{\sqrt{10^2+19^2}} & \frac{1}{19} \\ \frac{1}{19} & \frac{1}{\sqrt{10^2+19^2}} & \frac{1}{\sqrt{19^2+20^2}} & 0 & \frac{1}{10} & \frac{1}{20} \\ \frac{1}{\sqrt{10^2+19^2}} & \frac{1}{19} & \frac{1}{\sqrt{10^2+19^2}} & \frac{1}{10} & 0 & \frac{1}{10} \\ \frac{1}{\sqrt{19^2+20^2}} & \frac{1}{\sqrt{10^2+19^2}} & \frac{1}{19} & \frac{1}{20} & \frac{1}{10} & 0 \end{bmatrix} \tag{3-10}$$

STGCSEN-FixA2 中的邻接矩阵值为指数距离衰减邻接矩阵值，具体如下：

$$\boldsymbol{A}_{\text{FixA2}}=\begin{bmatrix} 0 & e^{-(0.1^2)} & e^{-(0.2^2)} & e^{-(0.19^2)} & e^{-(0.1^2+0.19^2)} & e^{-(0.19^2+0.2^2)} \\ e^{-(0.1^2)} & 0 & e^{-(0.1^2)} & e^{-(0.1^2+0.19^2)} & e^{-(0.1^2)} & e^{-(0.1^2+0.19^2)} \\ e^{-(0.2^2)} & e^{-(0.1^2)} & 0 & e^{-(0.19^2+0.2^2)} & e^{-(0.1^2+0.19^2)} & e^{-(0.19^2)} \\ e^{-(0.19^2)} & e^{-(0.1^2+0.19^2)} & e^{-(0.19^2+0.2^2)} & 0 & e^{-(0.1^2)} & e^{-(0.2^2)} \\ e^{-(0.1^2+0.19^2)} & e^{-(0.1^2)} & e^{-(0.1^2+0.19^2)} & e^{-(0.1^2)} & 0 & e^{-(0.1^2)} \\ e^{-(0.19^2+0.2^2)} & e^{-(0.1^2+0.19^2)} & e^{-(0.19^2)} & e^{-(0.2^2)} & e^{-(0.1^2)} & 0 \end{bmatrix} \tag{3-11}$$

STGCSEN-FixA3 中的邻接矩阵值为反对称衰减矩阵值，具体如下：

$$\boldsymbol{A}_{\text{FixA3}}=\begin{bmatrix} 0 & 1 & \frac{1}{2} & \frac{10}{19} & \frac{10}{\sqrt{10^2+19^2}} & \frac{10}{\sqrt{19^2+20^2}} \\ 1 & 0 & 1 & \frac{10}{\sqrt{10^2+19^2}} & \frac{10}{19} & \frac{10}{\sqrt{10^2+19^2}} \\ 2 & 1 & 0 & \frac{10}{\sqrt{19^2+20^2}} & \frac{10}{\sqrt{10^2+19^2}} & \frac{10}{19} \\ \frac{19}{10} & \frac{\sqrt{10^2+19^2}}{10} & \frac{\sqrt{19^2+20^2}}{10} & 0 & 1 & \frac{1}{2} \\ \frac{\sqrt{10^2+19^2}}{10} & \frac{19}{10} & \frac{\sqrt{10^2+19^2}}{10} & 1 & 0 & 1 \\ \frac{\sqrt{19^2+20^2}}{10} & \frac{\sqrt{10^2+19^2}}{10} & \frac{19}{10} & 2 & 1 & 0 \end{bmatrix} \tag{3-12}$$

其中， STGCSEN-DymA1 的邻接矩阵初值为全 1，STGCSEN-DymA2 的邻接矩阵的初值为 STGCSEN-FixA1 对应的固定邻接矩阵值，STGCSEN-DymA3 的邻接矩阵的初值为 STGCSEN-FixA2 对应的固定邻接矩阵值，STGCSEN-DymA4 的邻接矩阵的初值为 STGCSEN-FixA3 对应的固定邻接矩阵值。

不同邻接矩阵的 STGCSEN 模型在不同噪声场景下的 PESQ 和 STOI 分值分别如表 3-4 和表 3-5 所示。使用不同的邻接矩阵方案可以在不同场景下取得更优的 PESQ 和 STOI 分值，达到更好的语音增强效果。具体而言，使用动态邻接矩阵相比于使用固定邻接矩阵能够普遍达到更好的语音增强效果，无论是 STOI 和 PESQ 指标，四种动态邻接矩阵 STGCSEN 的表现普遍优于三种固定邻接矩阵的 STGCSEN，对于不同的固定邻接矩阵或动态邻接矩阵的不同初值设置上，与阵

列距离呈反比关系表征网络对不同麦克风信号幅度衰减补偿的邻接矩阵 $\boldsymbol{A}_{\text{FixA1}}$ 可以取得更好的语音增强效果，动态邻接矩阵中以 $\boldsymbol{A}_{\text{FixA1}}$ 为初值的 STGCSEN 网络也达到了最优的 PESQ 分值。

表 3-4 不同的邻接矩阵方案在四种噪声环境下的 PESQ 对比

Table 3-4 PESQ comparison of different adjacency matrix schemes in four noise environments

Model	BUS	CAF	PED	STR	Avg.
Noisy (channel 1)	1.28	1.15	1.18	1.26	1.22
STGCSEN	2.055±0.363	1.893±0.304	1.883±0.326	2.041±0.356	1.969±0.347
STGCSEN-FixA1	2.047±0.339	1.876±0.293	1.876±0.316	2.038±0.352	1.963±0.337
STGCSEN-FixA2	1.969±0.353	1.824±0.292	1.827±0.308	1.968±0.344	1.899±0.331
STGCSEN-FixA3	2.040±0.353	1.869±0.302	1.862±0.321	2.023±0.354	1.951±0.344
STGCSEN-DymA1	2.073±0.329	1.905±0.294	1.899±0.314	2.061±0.332	1.986±0.328
STGCSEN-DymA2	**2.078±0.330**	**1.911±0.291**	**1.908±0.317**	**2.067±0.334**	**1.991±0.329**
STGCSEN-DymA3	2.060±0.357	1.896±0.297	1.885±0.320	2.047±0.350	1.977±0.344
STGCSEN-DymA4	2.057±0.351	1.889±0.296	1.879±0.323	2.042±0.351	1.971±0.341

表 3-5 不同的邻接矩阵方案在四种噪声环境下的 STOI(%)对比

Table 3-5 STOI(%) comparison of different adjacency matrix schemes in four noise environments

Model	BUS	CAF	PED	STR	Avg.
Noisy (channel 1)	88.42	80.47	78.92	85.78	83.50
STGCSEN	92.56±5.37	**90.52±5.73**	89.61±6.17	**92.00±5.43**	**91.20±5.78**
STGCSEN-FixA1	92.37±5.37	90.30±5.78	89.30±6.25	91.76±5.52	90.96±5.87
STGCSEN-FixA2	91.54±6.19	89.67±6.23	88.79±6.61	91.08±5.97	90.33±6.30
STGCSEN-FixA3	92.46±5.36	90.21±5.80	89.29±6.20	91.84±5.42	90.99±5.84
STGCSEN-DymA1	92.56±5.33	90.45±5.76	89.61±6.19	91.99±5.34	91.18±5.80
STGCSEN-DymA2	**92.60±5.33**	90.42±5.74	**89.62±6.10**	91.97±5.36	91.17±5.77
STGCSEN-DymA3	92.55±5.32	90.47±5.71	89.54±6.16	91.96±5.44	91.18±5.79
STGCSEN-DymA4	92.49±5.38	90.36±5.78	89.47±6.16	91.95±5.38	91.11±5.80

3.3.3.4 多数据集对比实验

STGCSEN 在 CHiME3 数据集和 DNS-Set 数据集下的对含噪语音进行增强的

PESQ 和 STOI 指标如表 3-6 所示，在 CHiME3 数据集上，本文提出的 STGCSEN 在 PESQ 指标上对原含噪语音取得了约 52.9%的提升，在 STOI 指标上对原含噪语音取得了约 6.2%的提升；在 DNS-Set 数据集上，本文提出的 STGCSEN 在 PESQ 指标上对原含噪语音取得了约 16.2%的提升，在 STOI 指标上对原含噪语音取得了约 7.9%的提升。CHiME3 数据集包含了多种噪声场景，DNS-Set 数据集上包含不同类型、不同时长的室内混响类型，STGCSEN 在不同数据集上的 PESQ 和 STOI 指标均取得了明显的提升，说明本文提出的 STGCSEN 具有比较优异的算法稳健性及模型鲁棒性。

表 3-6 STGCSEN 在 CHiME3 和 DNS-Set 数据集下的 PESQ 和 STOI 对比结果

Table 3-6 PESQ and STOI comparison results of STGCSEN under CHiME3 and DNS-Set

	CHiME3		DNS-Set	
	Noisy (Channel 1)	STGCSEN	Noisy (Channel 1)	STGCSEN
PESQ	1.216	1.859	1.205	1.383
STOI	0.850	0.903	0.681	0.698

3.4 本章小结

本章构建了针对多通道语音增强任务的时空图卷积语音增强网络，且通过多维度的实验对比，证明了所设计算法的有效性和鲁棒性。本章分析了阵列信号的空间关联性和时间关联性并设计了空间信息提取模块和时频信息提取模块，分别实现了阵列信号的空间关联性提取和时频关联性提取；对于阵列信号的时-空相互关系进行分析并设计了时空信息融合模组，实现了对阵列信号时-空依赖性的融合提取；基于多个级联的 ST 模组和单独设计的通道信息融合模块完成了多通道时空图卷积语音增强网络的搭建，使用大量数据对所设计语音增强网络进行训练；从多个维度设计对比实验，多数据集对比试验结果表明本文设计的时空图卷积语音增强网络在多种噪声环境下均具有较好的适应性；多算法对比实验使用当前表现优异的集中多通道语音增强算法作为基准，PESQ 和 STOI 指标均有较大提升，其中 PESQ 指标提升 11%，且通过主观语音质量评价实验，从人们实际听觉感知的角度验证了本章所提算法在工程落地上的可行性。

本章的工作及部分实验结果(3.3.3.1, 3.3.3.2)已发表在本领域的旗舰国际会议 IEEE 2022 International Conference on Acoustics, Speech and Signal Processing (ICASSP)[53]上，部分干净语音、含噪语音、对比算法和本文所提出算法的增强语音片段可通过网站 https://ahuei.github.io/stgcsen 获取。

4 时空图卷积语音增强网络优化

为了进一步解决语音增强网络应用于小型智能人机交互设备所面临的瓶颈，进一步对时空图卷积语音增强网络进行优化设计，提升时空图卷积网络在多源突变场景下的语音增强性能，降低语音增强系统的软硬件资源占用率，本章从网络输入特征、网络参数量规模和增强语音的人耳感知三个维度设计了时空图卷积语音增强网络的优化策略。

针对时空图卷积语音增强网络构建过程中使用含噪语音信号幅度谱特征作为网络输入，未能充分利用含噪语音相位谱信息的问题，提出了基于复频谱的时空图卷积语音增强网络，提升了语音增强网络的性能上限；针对多ST级联的网络参数量、性能和时延问题设计了网络参数量优化方案，给出了不同场景下最优网络结构的选择方案，为算法在实际工程中的落地提供了技术支持；针对人耳对不同频带的语音有不同的感知质量的现实基础，设计了基于语音可懂度指数的损失函数，进一步提升了时空图卷积语音增强网络输出语音的可感知质量。

4.1 基于复频谱的时空图卷积语音增强网络构建

针对目标语音重建的相位信息损失问题，本节进行了基于复频谱的时空图卷积语音增强网络拓展。为了避免传统方法中只对目标分量的幅度谱或功率谱进行增强，而在重建目标语音阶段使用含噪语音相位谱从而引入相位噪声的问题，提出基于复频谱的时空图卷积语音增强网络，充分利用相位谱特征以达到同时对幅度谱和相位谱中的噪声信息进行抑制，从而有效提升语音增强系统的性能上限。本节在对复频谱与幅度谱进行分析的基础上，设计了遵循复数乘法规则的复频谱卷积模块，构建了基于复频谱的时空图卷积语音增强网络，并进行了多干扰源类型等复杂环境下的算法实验性能对比，证明了本节提出的基于复频谱的时空图卷积语音增强网络可以有效利用相位谱特征并实现复杂场景下更好的语音增强效果。

4.1.1 复频谱与幅度谱特征分析

基于时频域设计的语音增强算法，幅度谱[54]、功率谱[55]和梅尔功率谱[56]是比较常用的谱特征，将一段时间序列的语音信号通过 STFT 变换到时频域得到$\boldsymbol{\mathcal{X}}$，根据$\boldsymbol{\mathcal{X}}$获得对应的谱特征。在语音增强任务的重建目标语音阶段，通常将处理后的幅度谱、功率谱或梅尔功率谱与含噪语音的相位谱结合并进行短时傅里叶逆变

换(Inverse Short-Time Fourier Transform, ISTFT)，得到重建的时域目标语音。在相当长一段时间内，由于语音信号相位谱未表现出明确的特征，相位信息被认为是难以估计的，对含噪语音相位谱进行增强以获得更好的语音增强效果的研究一直寥寥。语音信号的频谱特征可以由其实部和虚部共同表征，也可以由幅度谱和相位谱共同表征，因此含噪语音信号的幅度谱和相位谱均包含干净语音信息及噪声信息，则以幅度谱等谱特征为输入，使用增强的幅度谱与含噪相位谱合成增强的语音始终包含了含噪相位谱携带的噪声信息，此类方法限制了语音增强问题的性能上限，无法达到理论上最优的语音增强效果。

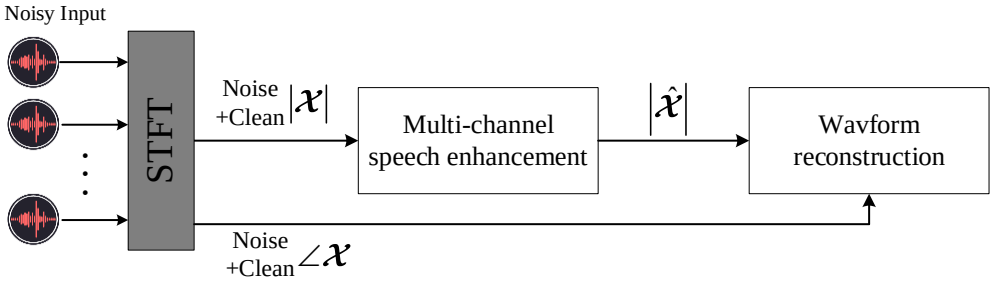


图 4-1 幅度谱语音增强重建语音过程中引入含噪相位谱噪声信息

Figure 4-1 Introducing noisy phase spectrum noise information in the process of speech enhancement with amplitude spectrum

Gerkmann 等针对语音增强中的相位问题进行了研究，认为在语音增强中相位的改善可以有效提升语音质量[57]，Wang 指出相位信息对于语音可感知质量是重要的，并提出一种基于复数谱映射的相位估计方法[58]。然而这些方法仅针对语音信号相位信息进行研究，或者将幅度谱和相位谱分别进行处理，如公式(4-1)所示，对幅度谱特征$|\mathcal{X}|$进行实数域的卷积获得估计的目标语音幅度谱特征，在将幅度谱特征与相位谱特征$e^{j\angle\mathcal{X}}$组合得到增强的语音信号，或者对幅度谱特征$|\mathcal{X}|$和相位谱特征$e^{j\angle\mathcal{X}}$分别进行实数域卷积，分别获得估计的幅度谱特征和估计的相位谱特征，然而在实际中，语音信号的幅度谱和相位谱是相互关联相互影响的，幅度谱特征和相位谱特征均与$\mathcal{X}$的实部和虚部相关联，对幅度谱的处理会影响到相位谱，反之亦然。

$$\mathcal{X}=\mathrm{Re}\{\mathcal{X}\}+j\,\mathrm{Im}\{\mathcal{X}\}=|\mathcal{X}|e^{j\angle\mathcal{X}} \tag{4-1}$$

另一方面，幅度谱和相位谱特征可以由信号的实部和虚部变换得到，因此将相位谱特征和幅度谱特征统一进行处理时可以转变为对信号实部和虚部的处理，以此达到充分利用语音信号相位谱特征，从而获得更好的语音增强效果。

本节分析了复频谱和幅度谱特征在携带信息上的差异，指出了仅使用幅度谱特征进行语音增强的算法性能上限的根源，得出对信号实部和虚部联合处理可以

达到充分利用语音信号相位谱特征的结论。

4.1.2 复频谱时空图卷积网络构建

不同于处理幅度谱特征时构造一个实数卷积核遵循实数卷积规则进行卷积，也不同于将含噪语音信号实部和虚部分为两个独立的部分分别进行训练，仅将处理后的实部和虚部进行组合得到增强的语音信号[59][60]，基于复频谱的时空图卷积网络的设计考虑语音信号幅度谱和相位谱之间的相互关联性，按照复数相乘的原理对复频谱的实部和虚部进行处理，即以语音信号复频谱的实部和虚部为输入，按照复数乘法规则设计复数域卷积核对复频谱特征进行复数域卷积，所设计的复数卷积核可以表达为：

$$\Theta = \mathrm{Re}\{\Theta\} + j\,\mathrm{Im}\{\Theta\} \tag{4-2}$$

则对 $\mathcal{X}$ 进行复数卷积的操作可表示为 $\mathcal{X} * \Theta$，具体表达如下：

$$\begin{aligned}&\mathcal{X} * \Theta = \\ &\left(\mathrm{Re}\{\mathcal{X}\} * \mathrm{Re}\{\Theta\} - \mathrm{Im}\{\mathcal{X}\} * \mathrm{Im}\{\Theta\}\right) + j\left(\mathrm{Re}\{\mathcal{X}\} * \mathrm{Im}\{\Theta\} + \mathrm{Im}\{\mathcal{X}\} * \mathrm{Re}\{\Theta\}\right)\end{aligned} \tag{4-3}$$

如图 4-2 所示，通过对复频谱进行复数卷积操作，可以有效利用语音信号中的相位谱特征，提升语音增强算法性能上限，为了加速复数卷积网络的收敛速度，在复数卷积之后添加了复数批归一化操作(Complex Batch Normalization, CBN)，CBN 之后采用 PReLU 激活函数增加复数卷积网络的非线性能力，所设计的复频谱卷积模块如下图所示，卷积过程中的步长设置为 $(2,1)$。

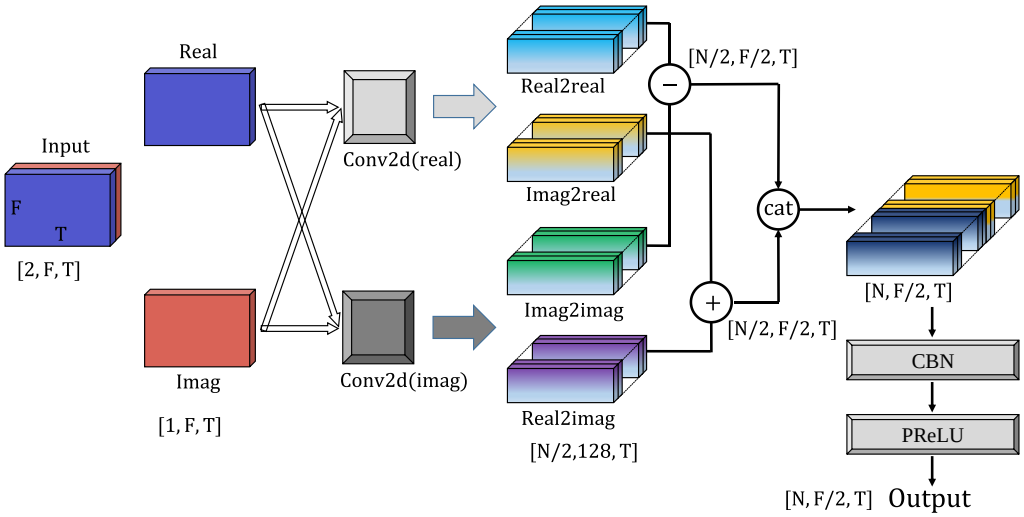


图 4-2 复频谱卷积模块结构

Figure 4-2 Complex spectral convolution module structure

本节结合复数相乘的规则，设计了以含噪语音信号实部、虚部为输入，实部

卷积核和虚部卷积核的复卷积核，将复频谱信号在复数域进行卷积的复频谱卷积模块，相比于以幅度谱为输入的实数域卷积，复频谱卷积可以对含噪语音相位谱特征进行处理，具有对含噪语音信号更丰富的表达能力，提升时空图卷积语音增强网络性能上限。

4.1.3 实验及结果分析

实验中数据集采用 CHiME3 数据集，模型训练参数及实验环境配置与 3.5 节相同。

表 4-1 复频谱时空图卷积网络及各对比算法的 PESQ

Table 4-1 PESQ scores of complex spectral spatiotemporal graph convolutional networks and comparison algorithms

Model	BUS	CAF	PED	STR	Avg.
Noisy (channel 1)	1.28	1.15	1.18	1.26	1.22
CNN	1.78±0.26	1.54±0.21	1.62±0.24	1.75±0.30	1.67±0.27
LSTM-IPD	1.91±0.06	1.49±0.20	1.51±0.26	1.73±0.34	1.66±0.35
UNet-GCN	1.66±0.20	1.44±0.17	1.49±0.19	1.64±0.26	1.56±0.18
STGCSEN	2.00±0.29	1.75±0.25	1.74±0.27	1.74±0.29	1.86±0.29
STGCSEN-complex	**2.13±0.36**	**1.93±0.31**	**1.93±0.35**	**2.12±0.37**	**2.03±0.36**

表 4-2 复频谱时空图卷积网络及各对比算法的 STOI

Table 4-2 STOI scores of complex spectral spatiotemporal graph convolutional networks and comparison algorithms

Model	BUS	CAF	PED	STR	Avg.
Noisy (channel 1)	0.91	0.81	0.81	0.88	0.85
CNN	0.92±0.03	0.89±0.04	0.89±0.04	0.91±0.03	0.90±0.04
LSTM-IPD	0.93±0.06	0.85±0.07	0.83±0.10	0.89±0.06	0.87±0.08
UNet-GCN	0.92±0.03	0.88±0.05	0.87±0.05	0.91±0.03	0.90±0.06
STGCSEN	0.92±0.06	0.89±0.06	0.88±0.06	0.92±0.06	0.90±0.06
STGCSEN-complex	**0.93±0.05**	**0.91±0.06**	**0.90±0.06**	**0.92±0.05**	**0.92±0.06**

表 4-1 和表 4-2 分别展示了使用复频谱网络和未使用复频谱网络的时空图卷积语音增强网络及其他几种基线模型在不同场景下的 PESQ 和 STOI 分值，根据表

格，使用复频谱时空图卷积网络在 BUS、CAF、PED、STR 四种场景下相比 STGCSEN 分别有 2.89%、4.00%、4.22%、8.40%的性能提升，综合指标有 3.78%的提升，在 STOI 各项指标上，相比 STGCSEN，STGCSEN-complex 在各噪声场景中也均有比较明显提升，表明使用复频谱卷积的方法有效利用了原含噪语音信号的相位信息，使得网络能同时对含噪语音的幅度谱和相位谱中的噪声信息进行抑制，获得增强的目标语音。

4.2 网络参数量优化

本节分析了时空图卷积语音增强框架中不同ST模组在语音增强性能、网络参数量及系统实时性的关系。根据实际应用场景不同，需要对时空图卷积语音增强算法的语音增强效果、网络参数量规模、系统增强语音的时延及软硬件系统资源调用等有不同的考量权重。首先比较了不同 ST 模组组成的 STGCSEN 在语音增强效果和参数量规模上的差异，提出在一般场景和边缘设备场景下最优的算法应用策略。其次分析了算法在增强语音过程中的时间延迟问题，给出了不同ST模组的最小时延。网络参数量和系统实时性的优化对算法的工程落地提供了技术支持。

4.2.1 面向实际应用的网络优化策略

智能语音交互具有场景突变、声源移动、多源混叠等多种复杂情况，不同的情况对语音增强算法在语音增强效果、计算资源占用、硬件资源调度、系统实时性及功耗等方面有不同的侧重。比如对于云端自动语音识别的语音增强前端模块，算法的语音增强性能往往是首要考量因素；对于高性能计算主机组成的多人远程实时会议系统，则主要关注算法的语音增强效果及系统实时性；而对于智能音箱、手环等小型边缘设备，在可接受的语音增强效果范围内，有效降低算法运行中的资源占用从而大幅降低系统功耗就显得尤为重要。网络参数量大小与算法运行过程中消耗的系统各项资源息息相关，本节首先比较了不同ST模组组成的时空图卷积语音增强网络在参数量规模-语音增强性能(PESQ)的关系，进一步设计了降低时空图卷积网络参数量的策略，给出了不同应用场景下适合的算法优化策略；另一方面，通过比较不同参数量的网络在语音增强过程中消耗的时间，给出了对系统实时性有较高要求的场景下的时空图卷积网络优化策略。

(1) 网络规模与语音增强性能的优化设计

本文设计的时空图卷积语音增强网络基于多个级联的ST模组，从网络本身的角度讲，增加级联的ST模组个数可以增加网络的深度从而使得网络具有更好的数

据拟合效果，使得网络获得更好的特征表达能力，获得更好的输出效果；另一方面，增加ST模组个数使得网络的参数量规模急剧增大，增加了网络的训练难度，训练时间，同时有可能使得网络训练过程难以收敛。本节横向比较了不同的ST模组个数组成的时空图卷积语音增强网络的语音增强性能及参数量规模，如下图所示：

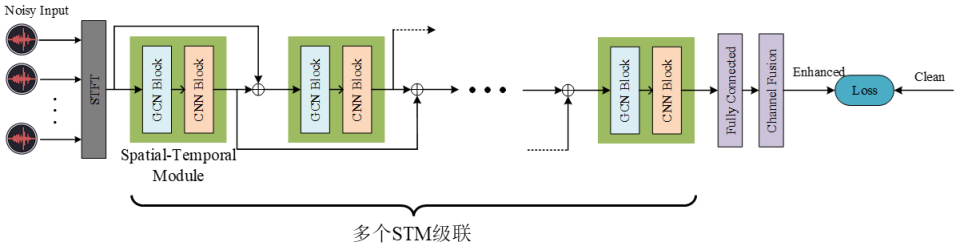
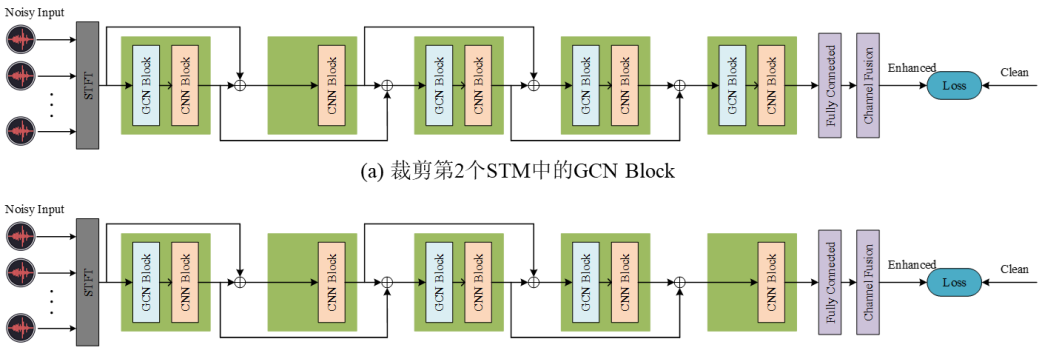


图 4-3 不同 ST 模组个数的时空图卷积语音增强网络结构

Figure 4-3 Spatial-temporal graph convolutional speech enhancement network structure with different number of ST modules

此外，针对 5 个 ST 模组组成的时空图卷积语音增强网络，通过对不同网络层的参数量分析，发现 GCN Block 的参数量规模大于 CNN Block，尝试对 GCN Block 的数量进行裁剪，设计如下两种新的时空图卷积语音增强网络结构：



(a) 裁剪第2个STM中的GCN Block

(b) 裁剪第2、5个STM中的GCN Block

图 4-4 裁剪 STM 中的 GCN Block 网络结构

Figure 4-4 Cut the network structure of GCN Block in STM

(2) 不同参数量的语音增强网络在系统时延中的优化设计

语音增强算法在处理信号时都会产生一些时间延迟。生理上，当系统的时延大于 300 ms 时，人耳已能感受到声音具有明显的卡顿，系统的输出已经能对人的主观感受造成严重的影响。因此对于视频会议、现场语音识别等场景中需要进行语音增强步骤时，需要对系统输入输出消耗时间进行严格控制

本文所设计时空图卷积语音增强算法的系统时延由算法处理语音信号所需最少采样点时间间隔和算法处理该段语音信号消耗时间两部分组成，如公式(4-4)所

示：

$$\tau_{total} = \tau_{sample} + \tau_{consume} \tag{4-4}$$

对τ_{sample}的分析需要从网络结构本身入手，时空图卷积网络中GCN Block对多个通道的相同时间帧的不同频点进行处理，每个时间帧的时间跨度为 16 ms，而前两个CNN Block中卷积核沿时间维度大小为5，步长为2，因此需要至少6个时间帧跨度，第三个 CNN Block 中卷积核沿时间维度大小为 3，步长为 1，则整个STGCSEN网络共需要 7 个时间帧跨度，即τ_{sample}为 112 ms。对$\tau_{consume}$的分析主要判断算法处理数据的时间消耗，可以通过对大量语音数据进行增强，计算得出平均每一个时间帧的时间消耗。

4.2.2 实验及结果分析

实验中数据集采用 CHiME3 数据集，模型训练参数及实验环境配置与 3.5 节相同，在网络规模与语音增强性能的优化设计阶段，比较了 4、5、6、7、8 个 ST 模组组成的 STGCSEN 网络在语音增强性能和参数量规模上的关系，分别命名为STGCSEN-4STM、STGCSEN-5STM、STGCSEN-6STM、STGCSEN-7STM、STGCSEN-8STM；同时，根据不同的网络结构，分析网络所需最少的τ_{sample}，训练好的各网络对全部测试集进行测试，并根据测试耗时及测试集长度求得$\tau_{consume}$，最终获得不同参数量的语音增强网络在语音增强效果、参数量规模及耗时三个指标上的比较结果。在处理耗时$\tau_{consume}$的计算上，本文以不同 ST 模组数量的STGCSEN网络处理 2256s 的音频所消耗的总时间为基准，求得 112 ms 音频段的处理耗时。所有对比实验使用相同的计算设备，CPU 为 Intel Xeon Gold 5218，GPU为Nvidia GeForce RTX 3090。

表 4-3 不同网络结构下参数量、耗时及语音增强性能比较结果

Table 4-3 Comparison results of parameter amount, total time consumption and speech enhancement performance under different network structures

Model	参数量(M)	最低采样耗时(ms)	处理耗时(ms)	PESQ	STOI
STGCSEN-4ST	22.2	96	4.23	1.773	0.900
STGCSEN-5ST	**30.6**	**96**	**4.17**	**1.830**	**0.903**
STGCSEN-6ST	39.0	112	4.62	1.780	0.902
STGCSEN-7ST	47.4	112	4.44	1.824	0.902
STGCSEN-8ST	55.9	128	6.14	1.777	0.900

不同网络结构下参数量及语音增强性能的对比结果如表 4-3 所示，随着 ST 模组数量增多，模型参数量也不断增加，但模型语音增强性能并不随着参数量增加而提升当 ST 模组数量大于 6 时，语音增强性能出现了下降的趋势，另外，在实际训练过程中，ST 模组数量越多，网络训练耗时越多，模型所需的最低采样耗时也逐渐增加，系统的实时性随 ST 模组数量增加而逐渐变弱。因此在实际应用中，设置 ST 数量为 5 即可满足大部分语音增强场景。但是在一些对系统实时性和模型参数量均有较高要求的场景(如手环等边缘设备)，且更多考虑系统功耗及计算资源消耗等因素而对语音增强的效果没有过高要求的情况下，可以选择 ST 模组数量为 4 的 STGCSEN 网络。

4.3 基于语音可懂度的损失函数构建

基于语音可懂度指数(Speech Intelligibility Index, SII)的损失函数，依据人耳对不同频带语音有不同权重的感知程度，使得网络训练过程中能够对重要语音频段分配更多权重。相比于传统最小均方误差损失函数(Mean Squared Error, MSE)，其在计算网络输出的估计目标语音信号与干净语音信号间差异时对每个频点权重相同，未体现出不同频率的音频对人耳感知影响的不同。通过构造拟合权重曲线为语音幅度谱特征中不同频点分配权重设计基于 SII 的损失函数，使得网络在训练过程中能够重点关注与人耳感知关系密切的频率范围，进一步提升输出语音的清晰度与可懂度。

4.3.1 声音频率与人耳听觉感知关系

表 4-4 语音可懂度权重与语音频带对应关系

Table 4-4 Correspondence between speech intelligibility weights and speech frequency band

频带	1	2	3	4	5	6	7	8	9
中心频率	160	200	250	315	400	500	630	800	1000
平均语音权重	0.0083	0.0095	0.0150	0.0289	0.0440	0.0578	0.0653	0.0711	0.0818
频带	10	11	12	13	14	15	16	17	18
中心频率	1250	1600	2000	2500	3150	4000	5000	6300	8000
平均语音权重	0.0844	0.0882	0.0898	0.0868	0.0844	0.0771	0.0527	0.0364	0.0185

人耳可以听到的声音频率范围是 20-20k Hz，且对不同频段声音的可感知程度不是均匀的。听觉机理中的一些研究表明，人耳对不同频率的语音质量(清晰度，可懂度)的主观感受是不同的，比如一段语音，其 1k-4k Hz 频段的内容对人耳感知语音的可懂度有更大的影响，同时，说话者说话内容的语音频谱，语言音素的变

化等都会影响到人耳对语音内容的感知。语音可懂度指数通过对每个频带分配频带重要性函数表示不同频段对人耳感知语音质量的影响程度，频带重要性函数表示每个频带对语音清晰度、可懂度的贡献。

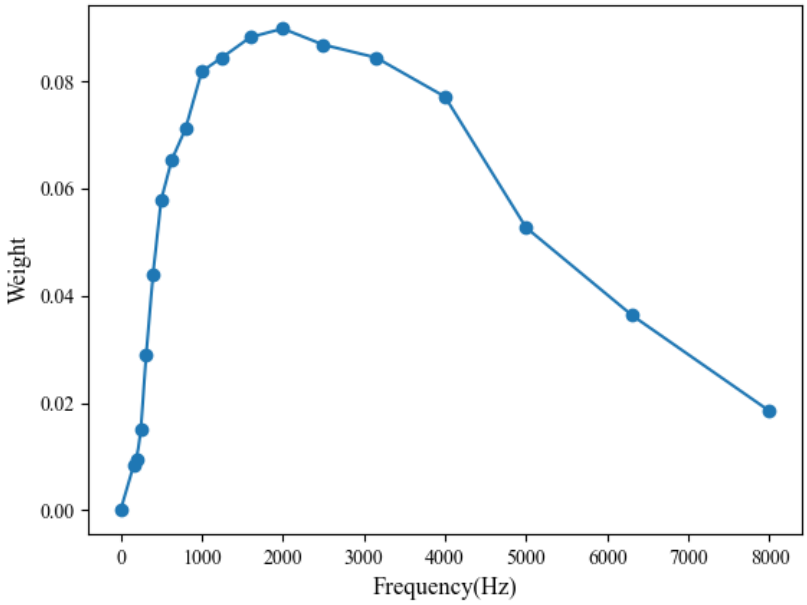


图 4-5 不同频率语音对人耳感知的重要性权重

Figure 4-5 The importance weight of different frequency speech to human ear perception

4.3.2 SII 损失函数设计

使用 MSE 计算生成的目标语音与干净语音的差异时相当于对每个频带分配的权重都相同，语音增强网络估计的时频点目标幅度谱特征为$\hat{\boldsymbol{Y}}_{f,t}$，其对应的时频点干净语音幅度谱特征$\boldsymbol{Y}_{f,t}$，公式表示如下：

$$\begin{aligned} Loss_{MSE} &= \frac{1}{\mathrm{F}\times\mathrm{T}}\sum_{f=1}^{\mathrm{F}}\sum_{t=1}^{\mathrm{T}}\left(\boldsymbol{Y}_{f,t}-\hat{\boldsymbol{Y}}_{f,t}\right)^2 \\ &= \frac{1}{\mathrm{F}}\times 1\times\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\left(\boldsymbol{Y}_{1,t}-\hat{\boldsymbol{Y}}_{1,t}\right)^2+\frac{1}{\mathrm{F}}\times 1\times\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\left(\boldsymbol{Y}_{2,t}-\hat{\boldsymbol{Y}}_{2,t}\right)^2+\cdots+\frac{1}{\mathrm{F}}\times 1\times\frac{1}{\mathrm{T}}\sum_{t=1}^{\mathrm{T}}\left(\boldsymbol{Y}_{F,t}-\hat{\boldsymbol{Y}}_{F,t}\right)^2 \end{aligned} \tag{4-5}$$

等式中的每一项表示每个频点的损失函数，数值 1 则代表 MSE 损失函数为每一个频点分配的权值相同。

对每个时频点在不同频段语音对人耳感知影响不同的基础上，对不同的频段按照语音可懂度指数分配不同的权重，设计基于语音可懂度权重的损失函数，从而达到更好的语音增强效果。网络输出的频点维度为 F，在本文的设计中，对 16k Hz 的采样信号进行窗长 512，滑动窗距离 256 个采样点的 STFT 变换，得到 F 为 257，其第 1 个维度表示直流分量，其余 256 个频点每个频点表示 31.25 Hz 的频

带，因此首先对表 4-4 中给出的 18 个中心频率-平均语音权重点的对应关系设计多项式拟合函数，公式表达如下：

$$\mathfrak{C}\left(f_{center}\right)=\sum_{k=0}^{\mathrm{K}}\alpha_k f_{center}^{\mathrm{K}-k} \tag{4-6}$$

其中，$\mathfrak{C}(\bullet)$ 表示多项式拟合函数，K 表示多项式拟合阶数，α_k 表示第 k 项多项式系数。根据多项式拟合函数获得第 f 个频点的权重，公式如下：

$$\omega_f=\mathfrak{C}\left(f_{center}=f\times\left(\frac{sr}{2}\times(\mathrm{F}-1)\right)\right) \tag{4-7}$$

对频点权重进行归一化：

$$\omega_f^{norm}=\frac{\mathrm{F}-1}{\sum_{f=1}^{\mathrm{F}}\omega_f}\omega_f \tag{4-8}$$

最终所设计的 SII 损失函数表达如下：

$$Loss_{SII}=\frac{1}{\mathrm{F}\times\mathrm{T}}\sum_{f=1}^{\mathrm{F}}\sum_{t=1}^{\mathrm{T}}\omega_f^{norm}\left(\boldsymbol{Y}_{f,t}-\hat{\boldsymbol{Y}}_{f,t}\right)^2 \tag{4-9}$$

SII 损失函数根据语音信号不同频带对人听觉的影响不同，将语音可懂度频带权重转换为谱特征不同频点权重，并将该权重归一化分配至不同频点，使得语音增强网络更有效地提取人耳重点关注的语音频带范围特征。

4.3.3 实验及结果分析

实验中数据集采用 CHiME3 数据集，模型训练参数及实验环境配置与 3.5 节相同，实验结果均取自训练 100 个 epoch 时保存的模型。

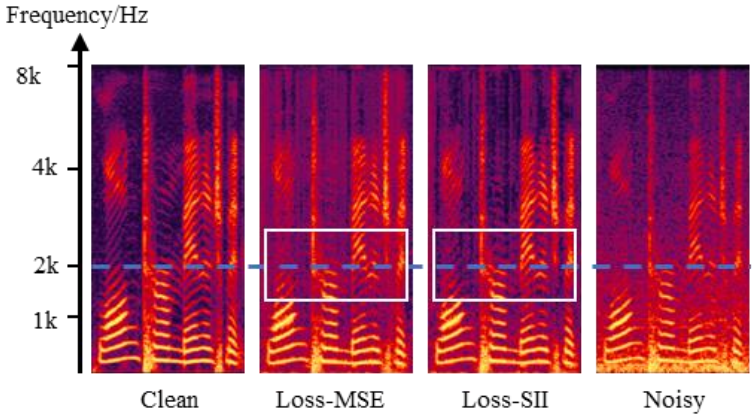


图 4-6 不同损失函数增强语音前后比较图

Figure 4-6 Comparison diagram before and after speech enhancement with different loss functions

根据 4.3.2 中语音可懂度权重拟合函数，有 51%的语音可懂度权重分布在频带

1k-3.15k Hz之间，26%的语音可懂度权重分布在1.6k-2.5k Hz之间，为了评估SII加权损失函数对语音清晰度恢复的有效性，我们计算了增强语音在1.6k-2.5k Hz人类听觉感知的最重要频段上的平均误差。

基于 SII 损失函数的网络在 BUS，CAF，PED，STR 四种不同噪声类型上将重要频段误差分别降低 1.8%，7.3%，5.9%，5.0%，证明了 SII 损失函数可以为语音中不同的频带分配不同的权重并重点关注对人主观感受影响更大的频带，达到提升语音可感知度的目标。

4.4 本章小结

本章分别从语音增强网络的复频域扩展、网络参数量规模优化和语音质量与人耳感知关系三个维度讨论了时空图卷积语音增强网络的优化设计方案。首先，分析了含噪语音信号幅度谱输入特征和复频谱输入特征在携带信息上的差异，论证了使用幅度谱特征作为输入时语音增强效果存在的性能上限，根据复数乘积原理以含噪信号频域实部和虚部为输入设计了基于复频谱的时空图卷积语音增强网络，有效利用含噪信号相位谱中的信息，进一步提升了语音增强的性能；其次，分析了不同ST模组数量构成的时空图卷积网络在语音增强效果、网络参数量及网络对含噪信号进行增强的耗时问题，给出了常规场景、实时场景和资源受限场景下最优的 ST 模组数量选择，对算法在不同场景下的实际工程落地提供了技术支持；最后，分析了人耳与语音频带的关系，不同频带的语音会对人耳造成不同的可感知程度，并根据语音可懂度指数设计了基于 SII 的损失函数，可以使网络在训练过程中对重要的频带分配更多的注意力，其输出的语音更符合人耳的听觉感知。本章所提出的三种针对时空图卷积语音增强网络的优化策略，可以进一步提升语音增强性能，保证算法精度的前提下降低对系统软硬件资源的消耗，使得网络输出的语音更符合人耳的听觉感知，具有较高的实际工程应用价值。

5 结论

5.1 本文工作总结

随着信息化、智能化进程的不断加快，各种各样的智能人机交互设备正快速融入人们生活的方方面面。例如，家庭场景中的智能音箱、沉浸式多媒体设备、语音监控系统等；在一些专业领域如养老医疗，行动不便的老年人通过语音指令与机器进行交互或借助机器与人更便捷的交互；工业场景中使用混合现实设备实现多人异地实时交互；尤其受疫情影响，通过视频会议方式协作交流变得更加频繁。所有这些智能语音交互场景都对语音增强系统的高效性和鲁棒性提出了更高的要求。而基于麦克风阵列的多通道语音增强技术是提升语音增强质量，提高语音识别精度和实现智能人机交互的关键技术。

然而，传统的基于波束成形的多通道语音增强算法由于在实际应用中缺乏准确的阵列和场景先验信息，语音增强性能往往无法达到最优，且在某些场景下甚至会出现严重的性能恶化。而目前基于机器学习的多通道语音增强算法没有结合声阵列信号的独特机理来进行网络模型的高效性、可控性和可解释性改造，尚无法适用于小型智能人机交互设备，且无法应对陌生多变场景。因此，本文拟针对声阵列信号中所隐含的复杂时空关联关系进行非欧空间的图理论建模，构建基于通道相关性和语音信号时频关联性的图聚合运算和动态邻接矩阵；搭建时空图卷积语音增强网络，在缺失阵列和场景准确先验信息的情况下，显著提升目标语音增强质量；提出复频谱网络拓展方法、模块参数量优化方法和基于语音可懂度的目标重建损失函数三种网络优化策略，进一步为算法在小型智能人机交互设备的落地应用提供可行方案和技术支持。具体包括：

完成了声阵列信号的图理论建模与设计，对多通道语音增强问题进行建模，得出对多通道信号进行增强实质是不同通道信号之间的加权叠加，将麦克风阵列构建为非欧氏空间中的图结构，多通道信号构建为图信号，设计了多通道信号的图卷积聚合运算，加权叠加中的权值由映射为图上邻接矩阵，完成了声阵列信号处理的图神经网络搭建。

针对多通道语音增强中声阵列信号的空间关联性、时间关联性和时-空依赖性，设计了空间信息提取模块、时频信息提取模块和时-空依赖性融合模组，有效提取声阵列信号特征，构建了基于时-空依赖性融合模组和通道信息融合模组的时空图卷积语音增强网络，有效提取目标语音、抑制噪声和干扰。实验结果证明本算法

在多种噪声环境、多种声源类型中均能在主客观评价指标中取得较好的效果。

针对目标语音重建的相位信息损失问题，进行了基于复频谱的时空图卷积语音增强网络拓展；进行了网络参数量和系统实时性的优化；进行了基于人耳听觉感知的语音可懂度网络训练损失函数设计。拓展的基于复频谱的时空图卷积语音增强网络提升了目标语音增强性能的上限，网络参数量和实时性优化为算法的工程落地提供了技术支持，基于语音可懂度的损失函数可使网络输出的重建目标语音更符合人耳听觉感知。实验结果为小型智能人机交互设备所面临的计算资源、系统功耗、信息失真等瓶颈提供了解决方案。

5.2 未来工作展望

本文提出的基于时空图卷积网络的多通道语音增强算法在多种变化噪声场景取得了比较优异的语音增强效果，但本算法仍然存在以下几点不足之处，需要在后续的研究中进一步改进：

(1) 本文所设计的 STGCSEN 框架以语音信号幅度谱或复频谱为输入特征，其时频关联性的提取是按照深度学习处理图像的思路来完成的，而语音信号有其自身固有独特特征，如音素、梅尔谱等，这些特征对于语音增强的效果提升也是非常有效的，后续需要考虑融合语音信号本身特征来进一步提升语音增强效果。

(2) 本文设计了时-空依赖性融合模组实现了声阵列信号的时空信息融合提取，但是从信号处理的顺序来看，该级联型结构仍是先提取空间信息再提取时频信息，对数据的异质性分析不足，后续应在此方向上进行更多的分析尝试。

(3) 本文使用的损失函数对语音质量的度量比较单一，即使采用了基于语音可懂度指数的损失函数获得了较好的语音听觉感知质量，但语音的评价维度非常多，清晰度、可懂度、舒适度等，后续应考虑从多个维度考量网络输出的语音质量并有针对性的设计语音增强网络。

参考文献

[1] 徐勇. 基于深层神经网络的语音增强方法研究[D]. 中国科学技术大学. 2015.

[2] Boll S. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1979, 27(2): 113-120.

[3] Ephraim Y, and Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1985, 33(2): 443-445.

[4] 王海艳. 基于统计模型的语音增强算法研究[D]. 吉林大学, 2011.

[5] Pandey A, and Wang D. A new framework for CNN-based speech enhancement in the time domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179-1188.

[6] Wang Y, Du J, Chai L, et al. A noise-aware memory-attention network architecture for regression-based speech enhancement[C]. //InterSpeech, 2020. 4501-4505.

[7] Tan K, and Wang D. A convolutional recurrent neural network for real-time speech enhancement[C]. //Interspeech, 2018. 3229-3233.

[8] Kim J, Khamy M, and Lee J. T-GSA: Transformer with Gaussian-weighted self-attention for speech enhancement[C]. //International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020. 6649-6653.

[9] Pascual S, Bonafonte A, and Serra J. SEGAN: Speech enhancement generative adversarial network[C]. //Interspeech, Stockholm, Sweden, 2017. 3642-3646.

[10] Mukaddim R, Weichmann A, and Varghese T. Photoacoustic delay-and-sum beamforming with spatiotemporal coherence factor[C]. //International Ultrasonics Symposium (IUS), 2020. 1-4.

[11] Capon J. High-resolution frequency-wavenumber spectrum analysis[J]. Proceedings of the IEEE, 1969, 57(8):1408-1418.

[12] Chaudhari K, Sutaone M, and Bartakke P. Adaptive diagonal loading of MVDR beamformer for sustainable performance in noisy conditions[C]. //IEEE Region 10 Symposium (TENSYMP), 2020. 1144-1147.

[13] Buckley K. Spatial/Spectral filtering with linearly constrained minimum variance beamformers[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 1987, 35(3):249-266.

[14] Wang J, and Mouthaan K. LCMV beamforming for conformal arrays using software defined radio[C]. //International Symposium on Antennas and Propagation (ISAP), 2021. 795-796.

[15] Chung H, Plourde E, and Champagne B. Basis compensation in non-negative matrix factorization model for speech enhancement[C]. //International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Shanghai, China, 2016. 2249-2253.

[16] Huemmer C, Astudillo R, and Kellermann W. An improved uncertainty decoding scheme with weighted samples for multi-channel DNN-HMM hybrid systems[C]. //Hands-free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 2017. 31-35.

[17] 奚琦. 基于盲源分离的自适应波束形成算法研究[D]. 北京交通大学, 2021.

[18] Wang D, and Chen J. Supervised speech separation based on deep learning: An overview[J].

IEEE Transactions on Acoustics, Speech, and Signal Processing, 2018, 26(10): 1702-1726.
[19] Ashutosh P, and Wang D. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain[C]. // International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 2019. 6875-6879.
[20] Bai S. Kolter J, and Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv, 2018, 1803.01271.
[21] Wang P, Chen P, and Yuan Y, et al. Understanding Convolution for Semantic Segmentation[C]. //Winter Conference on Applications of Computer Vision (WACV), 2018. 1451-1460.
[22] Zhao H, Zarar S, and Tashev I, et al. Convolutional-recurrent neural networks for speech enhancement[C]. //International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 2401-2405.
[23] Hao X, Shan C, and Xu Y, et al. An attention-based neural network approach for single channel speech enhancement[C]. //International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 6895-6899.
[24] Jiang Y, Wang D, Liu R, et al. Binaural classification for reverberant speech segregation using deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 2112-2121.
[25] Araki S, Hayashi T, Delcroix M, et al. Exploring multi-channel features for denoising-autoencoder-based speech enhancement[C]. //International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015. 116-120.
[26] Naohiro T, Tetsunori K, and Tetsuji O. Multi-channel speech enhancement using time-domain convolutional denoising autoencoder[C]. //Interspeech, Graz, Austria, 2019. 86-90.
[27] Minh H, Lee J, Lee B, et al. A cross-channel attention-based Wave-U-Net for multi-channel speech enhancement[C]. //Interspeech, 2020. 4049-4053.
[28] Fu Y, Wu J, Hu Y, et al. DESNet: A multi-channel network for simultaneous speech dereverberation, enhancement and separation[C]. //Spoken Language Technology Workshop (SLT), Shenzhen, China, 2021. 857-864
[29] Tzirakis P, Humar A, Donly J. Multi-channel speech enhancement using graph neural networks[C]. //International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021. 3415-3419.
[30] 杜功焕, 朱哲民, 龚秀芬. 声学基础[M]. 第二版. 南京: 南京大学出版社, 2001.
[31] Yu J, Xv J, Yu F. Geometry Design Based on Microphone Density Analysis[J]. Wireless Personal Communications, 2018, 103: 773-784.
[32] Wu Z, Pan S, Long G, et al. Graph wavenet for deep spatial-temporal graph modeling[C]. //International Joint Conference on Artificial Intelligence (IJCAI), 2019. 1907-1913.
[33] Li J, Han Z, Cheng H, et al. Predicting path failure in time-evolving graphs[C]. //Knowledge Discovery and Data Mining (KDD), 2019. 1279-1289.
[34] Yan S, Xiong Y, and Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]. //Association for the Advancement of Artificial Intelligence (AAAI), 2018. 7444-7452.
[35] Bruna J, Zaremba W, and Szlam A, et al. Spectral networks and locally connected networks on graphs[C]. //International Conference on Learning Representations (ICLR), 2014.

基于DNN方式的多通道语音增

[36] Defferrard M, Bresson X, and Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering[C]. //Conference and Workshop on Neural Information Processing Systems (NIPS), 2016. 3837-3845.

[37] Kipf T and Welling M. Semi-supervised classification with graph convolutional networks[C]. //International Conference on Learning Representations (ICLR), 2017.

[38] 吕波, 周杰. 线性天线阵列空间相关性函数的分析[J]. 南京信息工程大学学报(自然科学版), 2010, 2(05): 405-409.

[39] Hazen T, Shen W, and White C. Query-by-example spoken term detection using phonetic posteriorgram templates[C]. //IEEE Workshop on Automatic Speech Recognition & Understanding (WASRU), 2009. 421-426.

[40] Song C, Lin Y, Guo S, et al. Spatial-temporal synchronous graph convolutional networks: a new framework for spatial-temporal network data forecasting[C]. //Association for the Advancement of Artificial Intelligence (AAAI), 2020. 914-921.

[41] Barker J, Marxer R, Vincent E, et al. The third chime speech separation and recognition challenge: Dataset, task and baselines[C]. //IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015. 504-511.

[42] Harishchandra D., Vishak G, Ross C, et al. ICASSP 2022 deep noise suppression challenge[C]. //International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2022.

[43] Anonymous. LibriVox: Free Public Domain Audiobooks[J]. Reference Reviews, 2014, 28(1): 7-8.

[44] Gemmeke J, Ellis D, Freedman D, et al. Audio set: An otology and hman-lbeled dataset for audio events[C]. //International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2017. 776-780.

[45] Fonseca E, Pons J, Favory X, et al. Freesound datasets: A platform for the creation of open audio datasets[C]. //International Conference on Music Information Retrieval (ISMIR), 2017. 486-493.

[46] Allen B, Berkley A. Image method for efficiently simulating small-room acoustics[J]. The Journal of the Acoustical Society of America, 1979, 65(4): 943–950.

[47] Scheibler R, Bezzam E, and Dokmanic I. Pyroomacoustics: A python package for audio room simulation and array processing algorithms[C]. //International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2018. 351-355.

[48] Rix A, Beerends J, Hollier M, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]. //International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2001. 749-752.

[49] Taal C, Hendriks R, Heusdens R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]. //International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010. 4214-4217.

[50] Schoeffler M, Bartoschek S, Stöter F, et al. webMUSHRA-A comprehensive framework for web-based listening tests[J]. Journal of open research software, 2018, 6(1): 1-8.

[51] Park S, Lee J. A fully convolutional neural network for speech enhancement[C]. //Interspeech, 2017. 1993-1997.

[52] Rao W, Fu Y, Xu Y, et al. INTERSPEECH 2021 conferencing speech challenge: towards far-field multi-channel speech enhancement for video conferencing[C]. //IEEE Automatic Speech

Recognition and Understanding Workshop (ASRU), 2022.
[53] Hao M, Yu J, Zhang L. Spatial-temporal graph convolution network for multichannel speech enhancement[C]. //International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.
[54] Tan K, and Wang D. A Convolutional recurrent neural network for real-time speech enhancement[C]. //Interspeech, 2018. 3229-3233.
[55] Du Z, Lei M, Han J, et al. PAN: Phoneme-aware Network for monaural speech enhancement[C]. //International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. 6634-6638.
[56] Du Z, Zhang X, and Han J. A Joint Framework of Denoising Autoencoder and Generative Vocoder for Monaural Speech Enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1493-1505.
[57] T Gerkmann, M Krawczyk, and R Rehr. Phase estimation in speech enhancement-Unimportant, important, or impossible?[C]. //Convention of Electrical and Electronics Engineers in Israel, 2012. 1-5.
[58] Tan K, and Wang D. Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 380-390.
[59] Chiheb T, Olexa B, Zhang Y, et al. Deep complex networks[C]. //International Conference on Learning Representations (ICLR), 2018.
[60] Hu Y, Liu Y, Lv S, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[C]. //Interspeech, 2020. 2472-2476.

附录 A

声音传播衰减函数推导

声音传播衰减指声波在实际媒质中传播会出现声波随着距离而逐渐衰减的现象，主要原因是媒质的粘滞、热传导及物质的微观过程引起的弛豫效应等，声音在媒质中的衰减遵从指数衰减规律，即

$$S_{\text{收}}\left(f;\boldsymbol{r}_s\right)=S\left(f;\boldsymbol{r}_s\right)\exp\left(-\alpha\left|\boldsymbol{r}_k-\boldsymbol{r}_c\right|\right) \tag{A-1}$$

其中 α 为衰减系数，由文献[30]可得

$$\alpha=\frac{2\pi^2 f^2}{\rho_0\mu^3}\left[\frac{4}{3}\eta^{'}+\chi\left(\frac{1}{C_{\text{v}}}-\frac{1}{C_p}\right)+\sum_{i=1}^{n}\frac{\eta_i^{''}}{1+\left(4\pi^2 f^2\right)\tau_i^2}\right] \tag{A-2}$$

式(A-2)中中括号内部的三项加数分别代表粘滞衰减系数、热传导衰减系数和多种弛豫过程的衰减系数。

当各媒质影响因子都确定时，将所有影响因子由常数 ρ 代替，即

$$\alpha=\frac{2\pi^2 f^2}{\rho_0\mu^3}\left[\frac{4}{3}\eta^{'}+\chi\left(\frac{1}{C_{\text{v}}}-\frac{1}{C_p}\right)+\sum_{i=1}^{n}\frac{\eta_i^{''}}{1+\left(4\pi^2 f^2\right)\tau_i^2}\right]=\rho\frac{f^2}{\mu^3} \tag{A-3}$$

在 20°C 的干燥空气中，不考虑分子弛豫效应的影响，得到 ρ 值一般为 5.58×10^{-4} 。

附录 B

阵元空间相关性计算公式推导

根据高斯角能量分布模型，高斯角能量分布方程可以表示为：

$$p(\theta)=\frac{\kappa}{\sqrt{2\pi}\sigma}e^{-(\theta-\varphi)^2/2\sigma^2},\theta\in[-\pi+\varphi,\pi+\varphi] \tag{B-1}$$

其中，φ 表示中心到达角，σ 是角能量分布的标准差，κ 是标准化因子且

$$\kappa=\frac{1}{\mathrm{erf}\left(\frac{\pi}{\sqrt{2}\sigma}\right)},\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{B-2}$$

$\mathrm{erf}(x)$ 是误差函数且

$$\mathrm{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \tag{B-3}$$

在扩展角度较小时，κ 通常约为 1。

麦克风阵列中第 c 个阵元和第 c' 个阵元的空间相关性

$$\rho(i,j)=\int_{\boldsymbol{r}_s}u_i(\boldsymbol{r}_s)u_j(\boldsymbol{r}_s)^{\mathrm{H}}\,p(\boldsymbol{r}_s)d\boldsymbol{r}_s \tag{B-4}$$

假设角能量均匀分布，则 $\rho(i,j)$ 的实部和虚部分别表示如下：

$$\begin{aligned}\mathrm{Re}\left[\rho(i,j)\right]&=J_0(Z_l)+2\sum_{k=1}^{\infty}J_{2k}(Z_l)\cos(2k\varphi)\mathrm{sinc}(2k\Delta),\\ \mathrm{Im}\left[\rho(i,j)\right]&=2\sum_{k=1}^{\infty}J_{2k+1}(Z_l)\sin\left((2k+1)\varphi\right)\mathrm{sinc}\left((2k+1)\varphi\right)\end{aligned} \tag{B-5}$$

其中，$Z_l=2\pi\frac{(i-j)d}{\mu}$，$\Delta=\sqrt{3}\sigma$，$J_n(x)$ 是修正的第一类贝塞尔函数。

作者简历及攻读硕士学位期间取得的研究成果

一、作者简历

2019.09-2022.06 北京交通大学 电子科学与技术 工学硕士

2015.09-2019.06 北京交通大学 电子科学与技术 工学学士

二、发表论文

[1] **Hao M**, Yu J. Speech Enhancement Using Deep Complex Neural Network with Channel Attention[C]. //International Academic Exchange Conference on Science and Technology Innovation (IAECST), 2021. 530-534.

[2] **Hao M**, Yu J, Zhang L. Spatial-Temporal Graph Convolution Network for Multichannel Speech Enhancement[C]. //International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022. 6512-6516.

三、参与科研项目

［1］基于多模卷积的随机阵列结构化特征提取技术研究

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：郝明辉 签字日期：2022年 6月 2日

学位论文数据集

表 1.1： 数据集页

关键词*	密级*	中图分类号	UDC	论文资助
空间关联性；声阵列信号图聚合；时空信息融合；时空图卷积网络；多通道语音增强	公开			
学位授予单位名称*		学位授予单位代码*	学位类别*	学位级别*
北京交通大学		10004	工学	硕士
论文题名*		并列题名		论文语种*
基于时空图卷积网络的多通道语音增强算法研究				中文
作者姓名*	郝明辉		学号*	19120008
培养单位名称*		培养单位代码*	培养单位地址	邮编
北京交通大学		10004	北京市海淀区西直门外上园村 3 号	100044
学科专业*		研究方向*	学制*	学位授予年*
电子科学与技术		多维信号处理	三年	2022
论文提交日期*	2022 年 6 月			
导师姓名*	余晶晶		职称*	副教授
评阅人	答辩委员会主席*		答辩委员会成员	
	彭亚辉		黄琳琳；李居朋；李赵红；赵文山	
电子版论文提交格式 文本（√） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ） 推荐格式：application/msword；application/pdf				
电子版论文出版（发布）者		电子版论文出版（发布）地		权限声明
论文总页数*	63			
共 33 项，其中带*为必填数据，为 21 项。				