计算机科学
COMPUTER SCIENCE

基于GRU与自注意力网络的声源到达方向估计

何儒汉, 陈一帆, 余永升, 姜艾森

引用本文

何儒汉, 陈一帆, 余永升, 姜艾森. 基于GRU与自注意力网络的声源到达方向估计[J]. 计算机科学, 2023, 50(11A): 220900135-7.

HE Ruhan, CHEN Yifan, YU Yongshengand JIANG Aisen. Sound Source Arrival Direction Estimation Based on GRU and Self-attentive Network [J]. Computer Science, 2023, 50(11A): 220900135-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900135

基于 GRU 与自注意力网络的声源到达方向估计

何儒汉[1,2]　陈一帆[1,2]　余永升[3]　姜艾森[4]
1 纺织服装智能化湖北省工程研究中心　武汉 430200
2 武汉纺织大学计算机与人工智能学院　武汉 430200
3 武汉理工大学硅酸盐建筑材料国家重点实验室　武汉 430070
4 武汉纺织大学技术研究院　武汉 430200
(heruhan@wtu.edu.cn)

摘　要　基于神经网络的声源定位近年来受到广泛的关注,但如何缓解隐含 DOA 位置信息丢失、小样本数据等问题仍然是目前面临的挑战,因此提出了一种基于 GRU 和自注意力网络的声源到达方向估计方法。该方法采用对小型数据集效果较好的 GRU 作为骨干网络,弥补了纯净的声音数据采集困难的问题;同时,该方法使用多声道录音的声源形成训练集,经过短时傅里叶变换特征提取得到梅尔频谱图和声学强度矢量,进而形成由多通道语谱图以及归一化的主特征向量叠加的输入特征,避免了对语谱图与 GCC-PHAT 特征结合的隐式 DOA 信息的破坏,有效缓解了隐含 DOA 位置信息丢失问题;将其作为输入进入卷积循环神经网络模型进行监督学习获得模型参数。模型输出使用三维笛卡尔积坐标回归获得 DOA 位置估计,并增加自注意力网络在模型训练时进行参数回传,使得网络在训练的同时计算损失并预测关联矩阵,以解决预测定位和参考定位之间的最优分配。实验结果表明,该网络在不同混响条件和信噪比的环境下,均具有较高的定位准确率和鲁棒性。

关键词:声源到达方向估计;GRU;卷积神经网络;循环神经网络;自注意力
中图法分类号　TP391

Sound Source Arrival Direction Estimation Based on GRU and Self-attentive Network

HE Ruhan[1,2],CHEN Yifan[1,2],YU Yongsheng[3] and JIANG Aisen[4]
1 Hubei Provincial Engineering Research Center for Intelligent Textile and Fashion,Wuhan 430200,China
2 School of Computer Science and Artificial Intelligence,Wuhan Textile University,Wuhan 430200,China
3 State Key Laboratory of Silicate Materials for Architectures Wuhan University of Technology,Wuhan 430070,China
4 Science and Technology Institute,Wuhan Textile University,Wuhan 430200,China

Abstract　Neural network-based sound source localization has received wide attention in recent years. However,it is still challenging to mitigate the problems such as loss of implied DOA location information and small sample data. Therefore,a sound source arrival direction estimation method based on GRU and self-attentive network is proposed. The method uses GRU,which works well for small data sets,as the backbone network to compensate for the difficulty of pure sound data collection. At the same time,it uses sound sources from multichannel recordings to form a training set. After the short-time Fourier transform feature extraction to obtain the Meier spectrogram and acoustic intensity vector,then form the input features superimposed by the multi-channel speech spectrogram and the normalized main feature vector. Avoiding the implicit DOA information corrupted by the combination of speech spectrogram and GCC-PHAT features,effectively mitigating the loss of implicit DOA location information. It is used as input into the convolutional recurrent neural network model for supervised learning to obtain the model parameters. The model output uses 3D Cartesian product coordinate regression to obtain DOA location estimates,and adds a self-attentive network for parameter back-propagation during model training,enables the network to calculate the loss and predict the correlation matrix while training to solve the optimal allocation between predicted and reference localization. Experimental results show that the network has high localization accuracy and robustness under different reverberation conditions and signal-to-noise ratios.

Keywords　Sound source direction of arrival estimation,GRU,Convolutional neural network,Recurrent neural networks,Self-attention

近年来,声音事件检测往往与声源定位结合起来,共同解决一些现实中的问题,越来越多的人认识到声源定位的重要性。例如,在军事方面,通过定位狙击手周围的环境音,从而确定狙击手的位置,进行反狙击手活动;通过定位直升机发出

基金项目:国家自然科学基金面上项目(61170093)
This work was supported by the National Natural Science Foundation of China(61170093).
通信作者:陈一帆(2015363091@mail.wtu.edu.cn)

的噪音，对直升机进行定位等。在听力受损的人身上，声源定位可以帮助他们将声音可视化。机器人可以利用声源定位进行导航并与周围的环境互动。智能城市、智能家居和工业可以将其用于音频监控[1]。智能会议室可以在其他活动中识别语音，并使用这些信息来波束成形和增强语音，以召开电话会议或进行稳健的自动语音识别[2]。深度学习在图像识别、人工智能以及虚拟现实等领域取得了飞速的发展，而基于神经网络的麦克风阵列数据的声源表征研究却寥寥无几。现有的研究大多是关于医学超声影像方面的研究。Reiter 等采用 Alex-net 结构证明了根据传播波阵面的光声图像来可靠估计点源位置是可行的。Allman 等学者对该方法进行了更深入的研究，也有学者在传统的目标检测方法中加入了卷积神经网络结构来识别混响环境中的点声源。区别于传统目标检测对目标进行分类并确定其在图像中的位置这一主要目的，该方法可以在定位声源的同时，对光声图像中的真实声源和虚拟声源加以区分。

声源定位的任务是确定声源相对于麦克风的方向或位置，而只处理声音事件方向的估计则通常被称为声源到达方向(DirectionOfArrive，DOA)估计。目前有关声源定位的文献中所使用的 DOA 方法可以大致分为传统方法和基于深度神经网络(DNN)的方法。一些流行的传统方法包括基于到达时间差(TDOA)、导向响应功率(SRP)、多信号分类(MUSIC)以及通过旋转不变性技术(ESPRIT) 估计信号参数。这些方法在算法复杂性、阵列几何约束和声学场景的模型假设方面有所不同。其中像 MUSIC 这样的子空间方法可以应用于不同的阵列类型，并且可以产生多个源的高分辨率 DOA 估计。但传统方法在多个声源的环境下，很难估计其声源数量，并且对于混响和低信噪比的环境表现并不理想。

近年来，基于 DNN 的方法被用来克服传统方法的一些缺点，同时其在混响和低信噪比的情况下具有鲁棒性，如 SELD[3]，机器人可以在多扬声器场景中使用它进行基于声源的导航和自然交互[4-7]。目前基于 DNN 的方法在声源定位方向取得了不错的效果，在相同的混响场景中，得到了优于或与传统方法相当的性能。表 1 中列出了目前效果较好的基于 DNN 方法，大多数方法使用的是分类，从而在一组固定的角度估计声源存在的可能性。

表 1 基于 DNN 的声源到达方向估计方法

Table 1 DNN-based sound source arrival direction estimation method

方法	输入特征	DNN
Chakrabarty et al.[8-9]	Phase spectrum	CNN
Yalta et al.[6]	Spectral power	CNN Resnet
Xiao et al.	GCC	FC
Chu et al.[7]	GCC	CNN
Ferguson et al.[10]	GCC，cepstrogram	CNN
Adavanne et al.[11]	Phase and magnitude spectrum	CRNN

目前，声源到达方向估计的研究中，使用的数据集音频格式一般分为 microphone array(MIC)和 first-order Ambisonics (FOA)两种录制格式。现如今基于深度学习的 DOA 估计方法对于这两种格式的特征提取，都存在一定的隐含 DOA 位置信息丢失问题，进而影响实验结果的准确性。本文使用的特征具有 FOA 和 MIC 格式的信号功率和源 DOA 之间的三时映射，它是由多通道对数幅度线性频率谱图组成，在声谱图的每个频段上叠加有空间协方差矩阵(CM)的主特征向量的归一化版本。主特征向量被归一化，使得它可以表示 FOA 格式的声道间强度差(IID)，或者 MIC 格式的声道间相位差(IPD)。空间协方差矩阵与声谱图存在直接的频段对应关系，相互结合并不会丢失位置信息。

声源定位的数据集采集较为困难，很多 DOA 估计模型训练使用的都是小规模数据集。针对这一问题，本文使用循环神经网络的变体 Gate Recurrent Unit(GRU)来代替网络模型中的 LSTM 网络，不仅降低了整体模型参数，同时，相比 LSTM，GRU 对于较少的数据量可以取得更好的效果。

为了提高模型的训练精度，在模型框架中加入自注意力网络进行参数回传。在模型训练期间，同时将预测 DOA 与参考 DOA 进行计算得到距离预测矩阵。加入自注意力网络对预测矩阵进行训练，得到关联矩阵，实现参考 DOA 与预测 DOA 的最优分配，以此提高模型的训练精度。

基于上述的研究背景，本文以卷积循环神经网络为基础结构，实现了对于声源 DOA 的预测，并通过一定的数据集验证了本模型的精度。

1 国内外研究现状

1.1 传统声源定位算法

传统声源定位方法直接根据阵列信号理论建立麦克风信号与声源位置之间的数学模型，主要有 3 种：基于时延估计(TDOA)的声源定位、基于可控波束形成的声源定位、基于高分辨率谱估计的声源定位。基于机器学习的声源定位方法，则是通过对大量已知数据的学习，训练出一个机器学习模型来描述声源位置和阵列信号特征之间的映射关系。

基于时延估计来进行声源定位，实现简单、响应迅速，是目前应用最广泛的声源定位技术。麦克风阵列中每两个麦克风之间存在的信号到达时间差与声源的方向存在函数对应关系，通过确定阵列中各麦克风之间的信号延迟时间，再根据阵列与声源之间的几何关系，就可求解出声源的方向或具体位置坐标[12]。基于互相关函数的时延估计声源定位算法计算量小，实现速度快，但是其对噪声和混响的鲁棒性不够理想。尽管有各种频域加权因子的引入，但其也只能在高信噪比和低混响环境中才有较高的定位精度，随着噪声的升高和混响的增强，这类算法的性能会急剧下降。

基于可控波束形成的方法是一种空间方位搜索法[13]。波束形成的实质是空域滤波，通过改变阵列的导向矢量调整波束的指向，在整个空间不断进行方位扫描，当波束指向与声源方向一致时，就可以获得最大的波束输出功率。

高分辨率谱估计技术是一种基于矩阵分解的子空间技术，它从向量空间的角度来描述信号之间的关系，将信号划分为不同的子空间，并在此基础上定义了阵列信号的空间谱，用于描述声源信号的空间方位特性，再基于空间谱的极值点搜索得到声源信号的方位[14]。基于高分辨率谱估计的声源定位方法最大的优点是可以实现很高的角度分辨率，但其中涉及的矩阵运算以及全局搜索使得算法的时间复杂度很高，用这种方法来处理宽带语音信号时运算量往往会成倍地增加。

1.2 基于神经网络的定位方法

在过去的几年中，基于神经网络的声源定位算法在数据

扩充、特性工程、模型体系结构和输出格式等方面产生了很多新的研究成果。在数据扩充方面，Mazzon 等[15]提出了一种通过交换 FOA 格式信道的空间增强方法。Wang 等[16]重点研究了几种数据增强方法来克服数据稀疏问题，他们提出了一种新颖的四阶段数据增强方法，用于基于 ResNet-Conformer 的声学建模，且探索了两种空间增强技术，即音频通道交换(ACS) 和多通道模拟(MCS)。ACS 和 MDS 专注于通过扩展到达方向(DOA) 表示来增强有限的训练数据，使得使用增强数据训练的声学模型对声源的定位变化具有鲁棒性。

在模型体系结构方面，2015 年 Hirvonen[17]将声源定位制定为分类任务，通过将空间中的音频方位进行划分，将 10° 划分为一类，再通过分类算法进行训练，但这种方法存在 10° 以内的误差。2018 年，Adavanne 等[18]做了一项开创性的工作，他们使用端到端卷积循环神经网络(CRNN)，联合检测声音事件并估计相应的声源到达方向。这种方法降低了分类的误差率，使分类的音频划分更加细致准确，但模式整体准确率不高。模型使分类更加精确，但降低了正确预测类别的准确率。Hu 等[19]采用 EINV2 与音轨输出格式、置换不变训练和软参数共享策略，来检测同一类但在不同位置的不同声音事件，实现同时检测多个声源位置。该模型使用轨迹输出，将不同的声音类别与轨迹相对应，将多分类任务变为二分类或三分类，增加了分类的准确性。同时使用置换不变训练，使损失函数不会受到其他轨迹的影响，进一步加强模型的准确性和鲁棒性。Nguyen 等[20-21]探索了一种被称为序列匹配网络(Sequence Matching Network，SMN)的混合方法，该方法首先分别预测声音时间类别和声源到达方向估计，然后使用双向循环单元(BiGRU)匹配预测的声音类别和声音到达方向输出序列。Shimada 等[22]使用一种被称为活动耦合笛卡尔到达方向(ACCDOA)的表示技术，并将一种新的卷积神经网络(CNN)体系结构 D3Net[23]纳入用于 SELD 的 CRNN。

本研究提出了新的模型，模型方法使用卷积循环神经网络作为主干网络，加入了自注意力模块。针对空间定位误差，我们将源检测项包含在损失中，提高了整体性能。通过自注意力模块集成跟踪启发的损失项避免排列错误。模型提供了一种端到端的训练方式，可以处理具有可变源数量的动态变化条件，适合于现实生活中的注释记录。模型使用的特征提取方法由多通道对数幅度线性频率谱图组成，有效地避免了广义互相关变换带来的隐含 DOA 位置信息丢失，使模型训练更具有鲁棒性。

2 本文方法

2.1 特征提取

本文使用的特征包含两个主要成分：多通道对数线性语谱图和归一化主特征向量。

2.1.1 多通道对数语谱图

在语音信号的处理中，特征提取是将语音信号变为语谱图，将语谱图上的数据作为输入特征。语谱图的横轴 x 为时间，纵轴 y 为频率，(x,y)对应的数值代表在时间 x 时频率 y 的幅值。本文中使用的多通道对数线性语谱图是由短时傅里叶变换而来。设 M 为麦克风数，L 为声源数。M 通道麦克风阵列进行短时傅里叶变换(STFT)信号得到语谱图，如式(1)所示：

$$\boldsymbol{X}(t,f)=\sum_{i=0}^{L}S_i(t,f)\boldsymbol{H}(f,\phi_i,\theta_i)+\boldsymbol{V}(t,f)\in\mathbb{C}^M \tag{1}$$

其中，t 和 f 分别为时间指标和频率指标；S_i 是第 i 个源信号；$\boldsymbol{H}(f,\phi_i,\theta_i)$为第 i 个源 DOA 对应的频域转向矢量，ϕ_i 和 θ_i 分别为方位角和仰角；$\boldsymbol{V}$ 为噪声矢量。对于移动源，$\varphi_i=\varphi_i(t)$和 $\theta_i=\theta_i(t)$是关于时间的函数。多通道对数语谱图由语谱图计算而来。

$$\boldsymbol{LINSPEC}(t,f)=\log(\|\boldsymbol{X}(t,f)\|^2)\in\mathbb{R}^{M\times T\times F} \tag{2}$$

其中，T 是时间帧的数量，F 是频率箱的数量。

2.1.2 归一化主特征向量

目前音频的录制格式主要有 FOA 和 MIC 两种，与 MIC 相比，目前对于 FOA 格式的算法更多，实验表明，FOA 格式作为输入的性能要略优于 MIC 格式，但 MIC 格式在日常实践中更为常见。

基于特征向量的 FOA 阵列强度向量：FOA 阵列有 4 个通道，方向信号在声道间强度差中编码。FOA 阵列的典型转向矢量可以定义为：

$$\boldsymbol{H}^{\mathrm{FOA}}(t,\phi,\theta)=\begin{bmatrix}\boldsymbol{H}_w(t,\phi,\theta)\\ \boldsymbol{H}_x(t,\phi,\theta)\\ \boldsymbol{H}_y(t,\phi,\theta)\\ H_z(t,\phi,\theta)\end{bmatrix}=\begin{bmatrix}1\\ \cos(\phi)\cos(\phi)\\ \sin(\phi)\cos(\phi)\\ \sin(\phi)\end{bmatrix}\in\mathbb{R}^4 \tag{3}$$

其中，$\varphi_i=\varphi_i(t)$和$\theta_i=\theta_i(t)$分别为声源相对于阵列的时变方位角和仰角。

我们可以从主特征向量中计算一个基于特征向量的强度向量(EIV)来近似代替$[HX,HY,HZ]^{\mathrm{T}}$。首先，通过 $\boldsymbol{U}$ 的第一个元素对它进行规范化，这个元素对应于全向通道，丢弃第一个元素以获得$\bar{\boldsymbol{U}}$，然后取$\bar{U}$的实部，将其归一化，得到单位范数基于特征向量的 FOA 阵列强度向量(***EIV*** $\tilde{\boldsymbol{U}}$)。**FOA** 格式的输入特征是由四通道光谱图与三通道 ***EIV*** $\tilde{\boldsymbol{U}}$叠加而形成的。

基于特征向量的 MIC 阵列相位向量：对于远场麦克风阵列，方向信号编码在 IPD 中。任意几何 M 通道远场阵列的导向矢量可以用 $\boldsymbol{HMIC}(t,f,\phi,\theta)\in CM$ 来表示，公式如下

$$\boldsymbol{H}_m^{\mathrm{MIC}}(t,f,\phi,\theta)=\exp(-\tau 2\pi f d_{1m}(\phi(t),\theta(t))/c) \tag{4}$$

其中，c 为声速，$d_{1m}(\phi(t),\theta(t))$是声源在 MTH 传声器和参考($m=1$)传声器之间的到达距离，单位为米。到达的距离为：

$$\boldsymbol{d}_{1m}(\phi(t),\theta(t))=(\zeta_1-\zeta_m)^{\mathrm{T}}\begin{bmatrix}\cos(\varphi(t))\cos(\theta(t))\\ \sin(\varphi(t))\cos(\theta(t))\\ \sin(\theta(t))\end{bmatrix}\in\mathbb{R} \tag{5}$$

其中，ζ_1 和 ζ_m 分别是参考麦克风和第 m 个麦克风的笛卡尔坐标。

$$\tau_{1m}(\phi(t),\theta(t))=d_{1m}(\phi(t),\theta(t))/c \tag{6}$$

式(6)是声源在第 m 个麦克风和参考麦克风之间传播的到达时间差(TDOA)。从主特征向量 $\bar{\boldsymbol{U}}$ 中可以计算基于特征向量的相位向量(EPV)来近似代替$[d_{12}\cdots d_{1M}]^{\mathrm{T}}$。首先，我们通过 $\bar{\boldsymbol{U}}$ 的第一个元素来归一化 $\bar{\boldsymbol{U}}$，并将该元素选为参考麦克风，然后丢弃第一个元素以获得$\bar{\boldsymbol{U}}$，我们获取 $\bar{\boldsymbol{U}}$ 的相位角并通

过 $2\pi f/c$ 进行归一化，以获得基于特征向量的 MIC 阵列相位向量（$\boldsymbol{EPV}\ \widetilde{\boldsymbol{U}}$）。麦克风格式输入特征是通过堆叠 M 通道频谱图和（$M-1$）通道 EPV 形成的。

2.2 模型结构

本文使用的框架如图 1 所示。我们通过将主特征向量归一化，对于 FOA 格式的音频，能更有效地表示声道间的强度差，对于 MIC 格式的音频，能表示声道间的相位差；再同多通道对数语谱图结合形成输入特征。该方法将连续谱图中的一系列特征作为输入，预测每个帧中所有活动的声音事件的空间位置，生成所有声音活动事件的声源到达方向（DOA）预测轨迹并将其作为输出。其中每个声音事件都与 3 个回归器相关联，3 个回归器分别估计麦克风周围单位球面上 DOA 的三维笛卡尔坐标 X,Y,Z，并将每个声音事件的 DOA 输出都限制在[0,1]之间。在网络的末端加入自注意力网络，该网络的输入由距离矩阵 $\boldsymbol{D}$ 构成，并输出关联矩阵 $\boldsymbol{A}$。通过关联矩阵 $\boldsymbol{A}$ 实现参考 DOA 与预测 DOA 估计之间的最优对应。

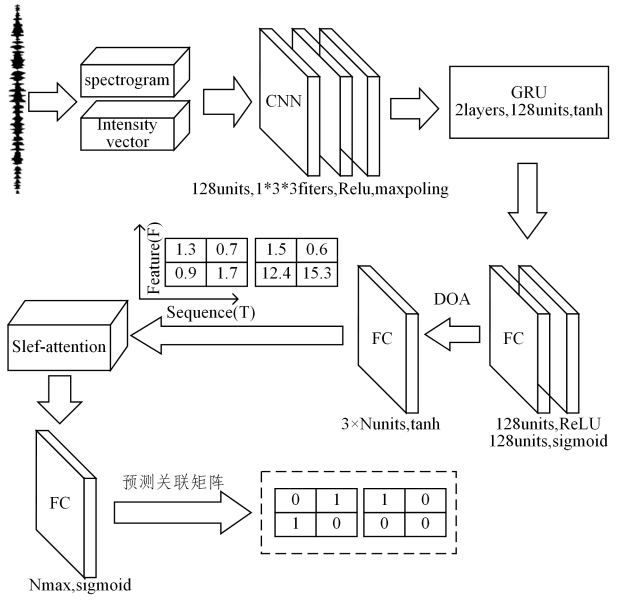


图 1 模型架构图

Fig. 1 Model architecture diagram

2.3 算法描述

算法 1 矩阵生成算法

输入：$\widetilde{\mathrm{X}}=[\widetilde{\mathrm{X}}_1(t),\cdots,\widetilde{\mathrm{X}}_i(t),\cdots,\widetilde{\mathrm{X}}_{M_t}(t)]$

$\widetilde{\mathrm{X}}=(\widetilde{x},\widetilde{y},\widetilde{z})$

$\mathrm{X}=[\mathrm{X}_1(t),\cdots,\mathrm{X}_j(t),\cdots,\mathrm{X}_{N_t}(t)]$

$\mathrm{X}=(x,y,z)$

输出：$\mathscr{D}_t=(M_t\times N_t)$

1. /* $\widetilde{\mathrm{X}}$为单个声源的预测空间位置，X 为单个声源的实际空间位置 x,y,z 为空间 3 个轴的坐标位置 */
2. for i ← 1 to M_t do

$\mathscr{D}_t[M_t][N_t]=\widetilde{\mathrm{X}}$

for i ← 1 to N_t do

$\mathscr{D}_t[M_t][N_t]=\mathrm{X}$

算法 2 估计算法

输入：$\mathscr{D}_t$

输出：$\mathscr{A}_t$

```
1. void forecast() //估计算法
  {
    for i->1 to nx //对于集合中的每一个节点
        if(从未匹配点 i 出发有增广路)
            匹配数++;
        输出 匹配数;
}
bool findpath(x)//寻找从 x 出发的对应项出的可增广路
{
//crossPath[x]=true;
for each edge(x,y) in G.E
    {
        if(y is not in crossPath)
        {
          add y into crossPath
          lastX=match[y];//lastX 是集合的上一个与 y 匹配的节点
          if(y is not matched or findpath(lastX))/* 如果 y 已经被了,那么试试从 lastX 能不能另外找到一条增广路,把当前增广路让给现在的 x */
          {
            match[y]=x;
            //match[x]=y,wrong !
            return true;//从 x 出发有增广路
              }
            }
}
return false;//从 x 出发没有增广路
}
```

算法 3 关联算法

输入：$\mathscr{D}_t,\mathscr{A}_t$

输出：LE_t

1. $\mathscr{A}_t=\mathscr{F}(\mathscr{D}_t)$ //$\mathscr{F}()$ 为估计算法，用于计算最小代价下的预测和参考最优关联
2. $\mathrm{K}_t\leftarrow\min(\mathrm{M}_t,\mathrm{N}_t)$
3. $\mathrm{LE}_t\leftarrow\frac{1}{\mathrm{K}_t}\sum_{ij}a_{i,j}(t)d_{ij}(t)=\frac{\|\mathscr{A}_t\odot\mathscr{D}_t\|_1}{\|\mathscr{A}_t\|_1}$ //LE_t为最佳误差

2.4 训练和测试数据集

该方法的输入是多通道音频，从每段音频中提取语谱图和主特征向量，共同输入进网络，结合成为输入特征。本文使用的数据集是四通道音频。网络训练使用的特征输入为 $7\times T\times 64$，每个特征都是由 4 通道对数线性语谱图和 3 通道的归一化主特征向量构成。T 是时间输入帧的数目，所有特征都使用 64 个 mel 波段进行训练。FOA 和 MIC 音频特征提取的不同之处在于主特征向量，FOA 的归一化主特征向量是基于特征向量的 FOA 阵列强度向量，MIC 的归一化主特征向量是基于特征向量的 MIC 阵列相位向量。

将上述特征输入神经网络，在如图 1 所示的神经网络中，CNN 主要是利用其多层网络训练时的局部位移不变特性。CNN 每一层都有 128 个 $3\times3\times7$ 维度的卷积核随着时间-频率-通道轴进行卷积，并通过激活函数（ReLU）。在每一层 CNN 之后，使用归一化函数来批量标准化输出，并使用 max-pooling 函数进行降维，从而保持序列长度不变。

最后一层 CNN 的输出通过激活函数，成为长度为 $128\times T/5\times8$ 的特征向量，并输入进循环神经网络（LSTM）。GRU

用于从 CNN 的输出中学习上下文信息，每层都有 128 个节点进行控制，并使用 tanh 激活函数进行激活。GRU 的网络模型简单，参数少，对于小规模的数据集来说训练速度更快。最后一层的 GRU 输出为 $T/5\times128$，之后是两个全连接层(FC)。第一个 FC 层使用 128 个节点进行线性激活，第二个 FC 层由 $3N$ 个节点组成，N 代表音频中的声音事件，每个音频事件由 3 个节点表示，分别对应声音事件位置的 X,Y,Z。为了保证 DOA 估计的每个轴的范围都在[−1,1]之间，使用 tanh 激活函数进行回归。

在网络训练期间，计算出相应声音事件的预测 DOA 和参考 DOA 的成对欧氏距离，形成距离矩阵 $\boldsymbol{D}$。将 $\boldsymbol{D}$ 作为输入特征输入自注意力网络进行训练，网络如图 2 所示，N 为可预测的活动事件的最大值。对于没有预测到的声音活动事件，使用远大于 1 的值进行填充，使其维度达到 $T/5\times N\times N$，N 为可预测的活动事件的最大值，矩阵的横轴为时间序列，纵轴为特征。将 D 通过 3 个 1×1 的卷积块，转换为 3 个向量，再将 $\boldsymbol{Q}$ 向量与 $\boldsymbol{K}$ 向量相乘，得到 attention map，将 attention map 与 $\boldsymbol{V}$ 向量相乘得到自注意力特征图。将自注意力特征图通过全连接层，输出关联矩阵 $\boldsymbol{A}$。我们对 $\boldsymbol{A}$ 矩阵提取其行最大值，继而获得所有声音事件活动的最大回归值。较高的值表示活动的概率最大，从而实现获得预测值和参考值的最优分配。

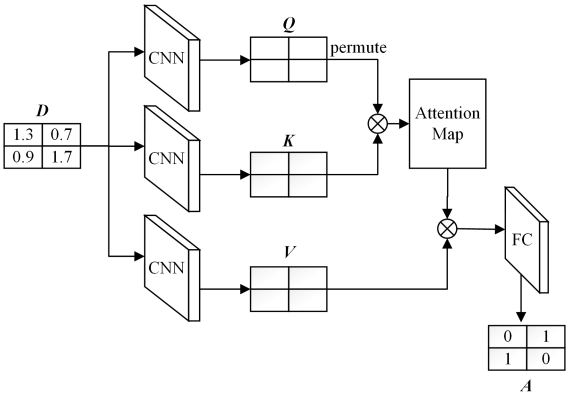


图 2　自注意力网络

Fig. 2　Self-attention network

2.5　模型优化与评价

在目前使用神经网络进行声源定位的模型框架中，输入特征都是由语谱图和相位频谱图构成。但语谱图与相位频谱图的结合会丢失相位频谱中隐含的 DOA 位置信息，特别是在多源的环境下，位置信息丢失更明显，所以本文使用强度向量与语谱图结合的方式，强度向量与语谱图存在直接的频率对应，所以两者结合并不会丢失位置信息。同时我们使用归一化的主特征向量近似代替单通道的强度向量，由三通道的归一化主特征向量代替三通道的强度向量，形成新的输入特征。

我们在网络模块中引入自注意力网络，在最后 FC 层的输出中通过笛卡尔坐标系转换，并于参考值形成距离矩阵，通过自注意力网络，使距离矩阵中的参考值和预测值形成最优的一一对应，并输出为预测矩阵。通过匹配其最优对应提高模型精度。

3　实验仿真及结果

3.1　数据集

本文使用的数据集是 TAU-NIGENS Spatial Sound Events 2020[24]，该数据集包含多个声音空间声场记录，集合了多种不同类别的声音，并记录了声音的方向和距离。我们在 TAU-NIGENS 2020 数据集上评估本文提出的方法，该数据集包含 FOA 格式和 MIC 格式两种音频，我们分别使用两种格式的音频对我们的方法进行评估。我们对 600 个时长一分钟的音频进行训练：500 个用于训练，100 个用于验证。并使用 200 个时长一分钟的音频进行测试。该数据集的采样率为 24 kHz，信噪比在 6 dB 到 30 dB 之间，800 个一分钟的音频中录制了大概 700 个声音事件，有 14 个声音类别，例如脚步声和狗叫等。

为了实验的鲁棒性，本文还使用了 TAU-NIGENS Spatial Sound Events 2021[25] 数据集，与 2020 数据集不同的是，2021 数据集中的声音时间分布更加自然，包含了一些额外的干扰音，使声源定位更加具有挑战性。

3.2　实验环境设置

实验在 liunx 操作系统下进行，硬件环境为 CPU Inter © Core™ i9-9900X，GPU 为 4 张 2080Ti 显卡，12GB 显存。软件环境为 python3. 8，pytorch1. 10，cudnn7. 6. 5，librosa0. 8. 1 等。使用短时傅里叶变换从音频中得到语谱图，并计算逐帧的空间特征向量，输出 npy 文件记录特征。

3.3　评估

为了评估本文提出的网络表现，我们使用以下两个指标进行评估：平均定位误差(LE)，即预测的位置与实际参考位置的误差平均值；定位召回率(LF)，即正确预测位置的概率(允许有 20°以内的误差)。

$$LE=\sum_{k}\theta_k \;/\; TP\;\theta_k k \tag{7}$$

$$LF=\frac{TP}{TP+FP} \tag{8}$$

其中，TP(truepositive)表示被模型正确预测为正值的正样本，FP(falsepositive)表示被模型预测为正值的负样本，TN(turenegative)表示被模型预测为负值的负样本，k 表示为未阈值的真阳性样本，θ 表示角度。同时将特征提取时间、模型训练时间、模型规模(kB)纳入评估范围进行对比。

3.4　实验结果与分析

实验在 TAU-NIGENS Spatial Sound Events 2020 上使用 CNN，GRU 和 FC 层的不同组合的各种架构，在多次组合后找到了最佳的模型组合。我们针对数据集使用不同的方式提取的空间信息，包括协方差矩阵(CM)和 RIR，实验表明 CM 的空间信息在与语谱图结合的过程中信息丢失比 RIR 更少。另一方面，DOA 估计的输出格式有两种可能：预测方位角和仰角，以及预测 DOA 在单位球面上的 x,y,z 坐标。为了确定这两种格式中的最佳输出格式，我们使用网络对两种格式的输出分别进行评估。在这个评估过程中，只有模型的输出权重参数在[1,5,50,500]的集合中被微调。在本研究中，我们选择默认 DOA 输出为方位角 180 °，仰角 60 °(数据集不包含这些 DOA 值的声音事件)，默认笛卡尔输出为 $x=0,y=0,z=0$。理论上，使用基于回归的 DOA 估计器优于基于分类的 DOA 估计器。基于回归的 DOA 估计器的优点

在于，网络不限于一组 DOA 角度，而是可以作为高分辨率连续 DOA 估计器来操作。最终我们的模型在 TAU-NIGENS Spatial Sound Events 2020 数据集上，LE 达到了 12.1°的误差，LF 达到了 80.17%的准确度。并且模型规模只有约 6 MB 的大小，一个 epoch 的训练时间是 96 s，与基线模型相比取得了不错的效果。

为了评估本文模型的效率，将其与 SELDnet，Wang 等提出的模型以及 Nguyen NTU 进行对比，比较了模型使用的模型规模(kB)、训练时间、验证速率，具体结果如表 2 所列。其中，实验的训练时间为每个推理步骤所用时长。实验结果表明，对比 SELDnet，本实验在推理时间相同的情况下，模型规模减少了 1 026 kB，训练时间减少了 11 s；对比 Wang 等提出的模型，模型规模减小了 157 726 kB，验证速率加快了 122.4 ms/step；相比 Nguyen NTU，模型规模减少了 103 455 kB，推理时间减少了 976 s，验证速率加快了 216 ms/step。表 2—表 4 在模型规模、模型在 FOA 音频格式上的精度，以及模型在 MIC 格式上的精度等几个方面进行比较。同时在 TAU2020 和 TAU2021 两个数据集上进行比较。如 Shimada 等提出的模型，只在 TAU2020 年的数据集上进行过训练，如 Zhang 等提出的模型，只在 TAU2021 上进行过训练。又如 Cao_Surrey 等提出的模型只能训练 FOA 格式音频。本文提出的模型兼具了上述图表的 4 种训练结果。

表 2 模型规模

Table 2 Model size

模型	训练时间/s	模型规模/kB	验证速率 ms/step
SELDnet	107	7140	531.1
Wang et al.[26]	1352	163840	523.5
Nguyen NTU	1072	109569	617.1
Ours	**96**	**6114**	**401.1**

表 3 FOA 格式模型准确率

Table 3 FOA format model accuracy

模型	TAU2020		TAU2021	
	LE	LF	LE	LF
SELDnet	22.8	60.1	24.1	43.9
Shimada et al.	10.2	79.1	—	—
Ye et al.	—	—	19.5	57.3
Yalta_HIT	—	—	20.1	71.1
Cao_Surrey	13.3	81.6	—	—
Ours	**9.7**	**82.06**	**17.5**	**72.4**

表 4 MIC 格式模型准确率

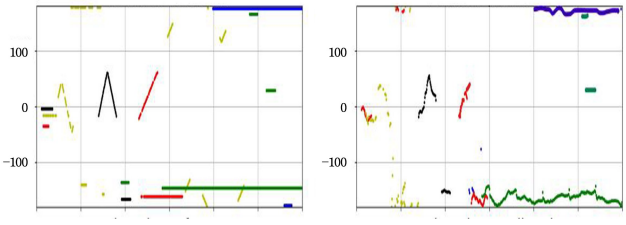
Table 4 MIC format model accuracy

模型	TAU2020		TAU2021	
	LE	LF	LE	LF
SELDnet	22.8	60.1	24.1	43.9
Politis_TAU	—	—	30.8	40.6
Phan_QMUL_	11.4	83.6	—	—
Bai_NWPU	—	—	66.5	35.5
Naranjo-Alcazar_UV	14.1	74.8	30.1	48.7
Song_LGE	13.3	73.0	—	—
Ours	**10.6**	**84.7**	**22.1**	**61.2**

Ye 等[27]的模型通过在 Transformer 解码器中添加基于跨模态注意(CMA)的注意层来估计输出，以便系统可以有效地学习到 DOA 位置信息。但该模型只能对 FOA 格式的音频进行训练，我们的模型不仅可以测试 FOA 格式，同时也对 MIC 格式的音频进行训练测试。

与 FOA 格式相比，MIC 格式录制的音频在通过深度学习模型训练时表现并不理想，而 Naranjo-Alcazar 等[28]着重研究了 MIC 音频格式的训练方式，通过改变卷积块实现在不修改基本框架的情况下获得更加稳健的系统，实现了 conv_standardpost 模块，并分析了不同的比值(ρ)对系统的影响。

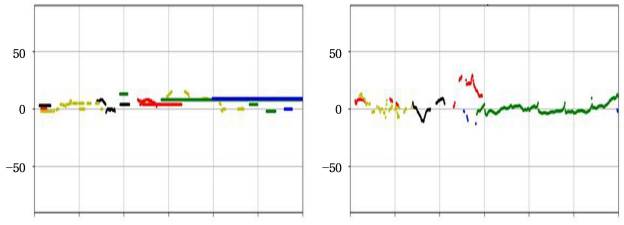
综上可知，本文提出的模型在保证训练速度快、模型规模小的同时，还可以达到不错的效果，并使用训练模型将结果可视化。模型采用 DOA 预测输出，图 3、图 4 中横坐标代替相同的时间帧，图 3 中纵轴代表方位，图 4 中纵轴表示距离。通过预测图和实际图的 DOA 位置对比，可以直观地感受到定位误差。虽然数据集获取和标注难度较大，但是本文提出的模型在实际声源定位的应用中仍有较大可行性。



(a) Azimuth reference (b) Azimuth predicted

图 3 DOA 位置方位对比

Fig. 3 DOA location and orientation comparison



(a) Elevation reference (b) Elevation predicted

图 4 DOA 位置距离对比

Fig. 4 DOA location and distance comparison

结束语 本文提出了一种新的端到端的模型方法，该模型的主干网络使用卷积循环神经网络，加入自注意力模块，可以处理可变源数、重叠场景的真实训练数据。在模型训练过程中，在相同的声学条件下，该模型实现了低定位误差，并改善了多个 DOA 回归量之间的跟踪性能，实现了参考值与测试值的最优分配。模型分别在 TAU-NIGENS Spatial Sound Events 2020 和 2021 年的数据集上，对 FOA 格式与 MIC 格式录制的音频文件进行训练测试，均取得了较好的效果。实验也证明了本文提出的模型在低信噪比和混响的场景下仍具有鲁棒性。模型对于双事件的音频定位具有不错的准确性，但对于多个事件的音频定位效果会差很多，无法做到同时定位多个音频事件，缺乏对于复杂场景下的鲁棒性。为此，后续会针对模型使用多轨迹同时训练与预测，使得模型可以同时预测多个声音事件，同时泛化不同环境的能力和多类别之间的检测能力，加强对于复杂场景下的训练与测试。

参 考 文 献

[1] HONG H, WANG M, FU M, et al. Sound Source Localization Sensor of Robot for Tdoa Method[C]// Third International Conference on Intelligent Human-machine Systems & Cybernetics. Zhejiang, China: IEEE, 2011: 19-22.

[2] SALVATI D，DRIOLI C，FORESTI G L. On the Use of Machine Learning in Microphone Array Beamforming for Far-Field Sound Source Localization[C]// 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing(MLSP). Vietrisul Mare，Italy：IEEE，2016：1-6.

[3] HIRVONEN T. Speech/Music Classification of Short Audio Segments[C]// IEEE International Symposium on Multimedia. IEEE，2015：2-9.

[4] TAKEDA R，KOMATANI K. Sound source localization based on deep neural networks with directional activate function exploiting phase information[C]// 2016 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2016：3-20.

[5] TAKEDA R，KOMATANI K. Discriminative multiple sound source localization based on deep neural networks using independent location model[C]// Spoken Language Technology Workshop. IEEE，2017.

[6] YALTA N，NAKADAI K，OGATA T. Sound source localization using deep learning models[J]. Journal of Robotics and Mechatronics，2017，29(1)：37-48.

[7] CHU F J，VELA P A. Deep grasp：Detection and localization of grasps with deep neural networks[J]. arXiv：1802. 00520，2018.

[8] CHAKRABARTY S，HABETS，EMANUËL A P. Multi-Speaker DOA Estimation Using Deep Convolutional Networks Trained with Noise Signals[J]. IEEE Journal of Selected Topics in Signal Processing，2019，13(1)：8-21.

[9] CHAKRABARTY S，HABETS E. Multi-speaker localization using convolutional neural network trained with noise[J]. arXiv：1712. 04276，2017.

[10] FERGUSON E L，WILLIAMS S B，JIN C T. Sound Source Localization in a Multipath Environment Using Convolutional Neural Networks[J]. arXiv：1710. 10948，2017.

[11] ADAVANNE S，POLITIS A，VIRTANEN T. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network[C]// 2018 26th European Signal Processing Conference(EUSIPCO). IEEE，2018：1462-1466.

[12] ZHOU Z，RUI Y，CAI X，et al. Constrained total least squares method using TDOA measurements for jointly estimating acoustic emission source and wave velocity[J]. Measurement，2021，182：109758.

[13] ZHANG Y D. Research on Microphone Array Sound Source Localization and Beamforming Fof Speech Interaction[D]. Xiamen：Xiamen University，2019.

[14] CHEN Y，HSU Y，BAI M R. Multi-channel end-to-end neural network for speech enhancement，source localization，and voice activity detection[J]. arXiv：2206. 09728，2022.

[15] MAZZON L，KOIZUMI Y，YASUDA M，et al. First order ambisonics domain spatial augmentation for DNN-based direction of arrival estimation[J]. arXiv：1910. 04388，2019.

[16] WANG Q，DU J，WU H X，et al. A four-stage data augmentation approach to ResNet-Conformer based acoustic modeling for sound event localization and detection[J]. arXiv：2101. 02919，2021.

[17] HIRVONEN T. Classification of spatial audio location and content using convolutional neural networks[C]// Audio Engineering Society Convention 138. Audio Engineering Society，2015.

[18] ADAVANNE S，POLITIS A，NIKUNEN J，et al. Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks：10. 1109/JSTSP. 2018. 2885636[P]. 2018.

[19] HU J，CAO Y，WU M，et al. Sound Event Localization and Detection for Real Spatial Sound Scenes：Event-Independent Network and Data Augmentation Chains[J]. arXiv：2209. 01802，2022.

[20] NGUYEN T N T，JONES D L，GAN W S. A sequence matching network for polyphonic sound event localization and detection[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP). IEEE，2020：71-75.

[21] NGUYEN T N T，NGUYEN N K，PHAN H，et al. A general network architecture for sound event localization and detection using transfer learning and recurrent neural network[C]// 2021 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP 2021). IEEE，2021：935-939.

[22] SHIMADA K，KOYAMA Y，TAKAHASHI N，et al. ACCDOA：Activity-coupled cartesian direction of arrival representation for sound event localization and detection[C]// 2021 IEEE International Conference on Acoustics，Speech and Signal Processing(ICASSP arXiv). IEEE，2021：915-919.

[23] TAKAHASHI N，MITSUFUJI Y. Densely connected multi-dilated convolutional networks for dense prediction tasks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021：993-1002.

[24] POLITIS A，ADAVANNE S，VIRTANEN T. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection[J]. arXiv：2006. 01919，2020.

[25] POLITIS A，ADAVANNE S，KRAUSE D，et al. A dataset of dynamic reverberant sound scenes with directional interferers for sound event localization and detection[J]. arXiv：2106. 06999，2021.

[26] WANG Q，WU H，JING Z，et al. The USTC-iFlytek system for sound event localization and detection of DCASE2020 challenge[J]. IEEE AASP Chall. Detect. Classif. Acoust. Scenes Events，2020，17(1)：5-13.

[27] YE Z，WANG X，LIU H，et al. Sound Event Detection Transformer：An Event-based End-to-End Model for Sound Event Detection[J]. arXiv：2110. 02011，2021.

[28] NARANJO-ALCAZAR J，PEREZ-CASTANOS S，FERRANDIS J，et al. Sound event localization and detection using squeeze-excitation residual CNNs[J]. arXiv：2006. 14436，2020.

HE Ruhan，born in 1974，Ph.D，professor，is a member of China Computer Federation. His main research interests include machine learning，computer vision and multimedia retrieval.

CHEN Yifan，born in 1999，postgraduate. His main research interests include machine learning and sound source localization.