DOI:10.11918/202104081

基于梅尔频谱分离和 LSCNet 的声学场景分类方法

费鸿博[1,2],吴伟官[1,2],李 平[1,2],曹 毅[1,2]

(1.江南大学 机械工程学院,江苏 无锡 214122;2.江苏省食品先进制造装备技术重点实验室(江南大学),江苏 无锡 214122)

摘 要: 针对现有频谱分离方法进行声学场景分类研究时其分类准确率不高的问题,提出了一种基于梅尔频谱分离和长距离自校正卷积神经网络(long-distance self-calibration convolutional neural network,LSCNet)的声学场景分类方法。首先,介绍了频谱的谐波打击源分离原理,提出了一种梅尔频谱分离算法,将梅尔频谱分离出谐波分量、打击源分量和残差分量;然后,结合自校正神经网络和残差增强机制,提出了一种长距离自校正卷积神经网络;该模型采用频域自校正算法以及长距离增强机制来保留特征图原始信息,通过残差增强机制和通道注意力增强机制加强了深层特征与浅层特征间的关联度,且结合多尺度特征融合模块,以进一步提取模型训练中输出层的有效信息,从而提高模型的分类准确率;最后,基于 Urbansound8K 和 ESC-50 数据集开展了声学场景分类实验。实验结果表明:梅尔频谱的残差分量能够针对性地减少背景噪音的影响,从而具有更好的分类性能,且 LSCNet 实现了对特征图中频域信息的关注,其最佳分类准确率分别达到 90.1% 和 88%,验证了该方法的有效性。

关键词: 声学场景分类;梅尔频谱分离算法;长距离自校正卷积神经网络;频域自校正算法;多尺度特征融合

中图分类号: TP391.42 **文献标志码:** A **文章编号:** 0367-6234(2022)05-0124-07

Acoustic scene classification method based on Mel-spectrogram separation and LSCNet

FEI Hongbo[1,2], WU Weiguan[1,2], LI Ping[1,2], CAO Yi[1,2]

(1. School of Mechanical Engineering, Jiangnan University, Wuxi 214122, Jiangsu, China; 2. Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment and Technology (Jiangnan University), Wuxi 214122, Jiangsu, China)

Abstract: When the existing spectrogram separation methods are used for acoustic scene classification research, the classification accuracy of these methods is not high. To solve the problem, an acoustic scene classification method based on Mel-spectrogram separation and long-distance self-calibration convolutional neural network (LSCNet) was proposed. Firstly, the working principles of spectrogram harmonic/percussive-source separation were presented. A Mel-spectrogram separation algorithm was proposed, which can separate the Mel-spectrogram into harmonic components, percussive source components, and residual components. Then, LSCNet was designed combining self-calibration convolutional network and residual enhancement mechanism. The model adopts frequency domain self-correction algorithm and long-distance enhancement mechanism to retain the original information of the feature map, strengthens the correlation between deep and shallow features through residual enhancement mechanism and channel attention enhancement mechanism, and combines multi-scale feature fusion module to further extract the effective information of the output layer in model training. Finally, acoustic scene classification experiments were conducted on Urbansound8K and ESC-50 datasets. Experimental results show that the Mel-spectrogram residual components (MSRC) could specifically reduce the influence of background noise, thereby indicating a better classification performance. The LSCNet could realize the attention to the frequency domain information in the feature map, and its best classification accuracy reached 90.1% and 88% respectively, which verified the effectiveness of the proposed method.

Keywords: acoustic scene classification; Mel-spectrogram separation algorithm; LSCNet; frequency domain self-calibration algorithm; multi-scale feature fusion

在自然环境中,声音是传递信息的重要媒介,也是人类听觉感知系统的重要组成部分。在对复杂环境中的声音事件进行感知方面,人类的固有能力使其不仅能同时捕捉多个声源的信息,且能有选择地屏蔽周围的背景噪音。当前,随着智能技术的快速发展,许多智能设备虽能高效地识别语音信息和声纹信息[1],但对复杂环境中的声音事件进行分类识别时,往往会因为背景噪音的影响而导致分类识别的准确率不高[2-3]。

收稿日期: 2021-04-19
基金项目: 高等学校学科创新引智计划(B18027);
江苏省"六大人才高峰"计划(ZBZZ-012);
江苏省优秀科技创新团队基金(2019SK07)
作者简介: 费鸿博(1997—),男,硕士研究生
通信作者: 曹 毅,caoyi@jiangnan.edu.cn

针对声学场景分类研究，国内外诸多学者大都采用特征融合、特征补偿等方法对特征图中的信息进行增强处理[3-8]。其中，文献[3]提出了一种将梅尔频谱、短时能量、声学似然特征和静音时间 4 个特征通过拼接以实现特征间融合的方法，该模型在 Dcase2017 数据集上的平均准确率达到了 71.1%；文献[4]提出了一种将梅尔倒谱系数和伽马通倒谱系数通过线性叠加处理进行融合的方法，并基于卷积神经网络开展了特征融合实验，其模型的分类准确率达到了 87.7%；文献[5]提出了一种梅尔倒谱系数和 Teager 能量算子混合的特征融合方法，在一定程度上起到了抑制背景噪声的作用，其在血管声学分类任务中精度达到了 99.5%；文献[6-8]采用 CNN 和 DCNN 作为分类模型开展了声学场景分类研究。

综上所述，针对声学场景分类研究中的特征处理方法虽然已经有较为深入的研究，但不难发现：1）已有研究成果中，并未对背景噪音进行针对性处理，这会导致模型将正确的音频类别错误地识别成背景噪音的音频类别或与背景噪音相近的其他音频类别，从而降低模型的分类准确率；2）现有声学模型相对单一，未能对特征图中的频域信息进行关注，导致了模型的分类准确率不高，泛化性能不强等问题。

基于上述问题，本文提出了一种基于梅尔频谱特征分离和长距离自校正卷积神经网络（long-distance self-calibration convolutional reural network，LSCNet）的声学场景分类方法。首先，基于谐波打击源分离原理提出了一种梅尔频谱分离算法，将梅尔频谱分离出谐波分量、打击源分量和残差分量，其中梅尔频谱的残差分量针对性地降低了背景噪音对模型的影响；然后，设计了一种长距离自校正神经网络，该模型采用频域自校正算法以及长距离增强机制，可在特征提取的同时保留特征图原始信息，并结合多尺度特征融合模块，以进一步保留模型训练中输出层的有效信息；最后，基于 LSCNet 模型利用 Urbansound8K 和 ESC50 数据集开展了声学场景分类实验，结果验证了所提方法及模型的有效性。

1　声学特征分离

1.1　频谱的谐波打击源分离

频谱的谐波打击源分离是将频谱中的混合信号分离出谐波分量和打击源分量。频谱是一种由时间和频率两个维度构成的图片，可通过分帧后的音频信号经过短时傅里叶变换转换而来。将中值滤波器应用于频谱分离即可获得频谱的谐波分量和打击源分量[9]，下面介绍其具体的工作原理。

首先，通过短时傅里叶变换将音频信号转化成频谱信号 $\boldsymbol{S}$，频谱的时域分量 $\boldsymbol{S}(t)=[t_1,t_2,\cdots,t_m]$ 和频域分量 $\boldsymbol{S}(f)=[f_1,f_2,\cdots,f_n]$，其中 m 为时间帧的数量，n 为频率帧的数量。

其次，定义中值滤波计算如下：

$$y(a)=\text{median}\{x(a-j,a+j),\ j=(l-1)/2\} \tag{1}$$

式中：$x(a)$ 为输入信号序列，l 为滤波器的长度，$y(a)$ 为输出信号序列，median{} 为中值滤波器函数。基于式(1)将频谱 $\boldsymbol{S}$ 的时域分量 $\boldsymbol{S}(t)$ 和频域分量 $\boldsymbol{S}(f)$ 的绝对值作为输入信号，则输出信号为频谱 $\boldsymbol{S}$ 的谐波增强部分 $\boldsymbol{H}$ 和打击源增强部分 $\boldsymbol{P}$：

$$\boldsymbol{P}=\text{median}\{|\boldsymbol{S}(t)|,l_{\text{P}}\} \tag{2}$$

$$\boldsymbol{H}=\text{median}\{|\boldsymbol{S}(f)|,l_{\text{H}}\} \tag{3}$$

式中：l_{P} 为打击源增强部分的滤波器长度，l_{H} 为谐波增强部分的滤波器长度。

然后，基于频谱的谐波增强部分和打击源增强部分，可计算出其各自的相对分量 $\boldsymbol{M}_{\text{H}}$ 和 $\boldsymbol{M}_{\text{P}}$ 如下：

$$\boldsymbol{M}_{\text{H}}=\frac{\boldsymbol{H}^p}{\boldsymbol{H}^p+\boldsymbol{P}^p} \tag{4}$$

$$\boldsymbol{M}_{\text{P}}=\frac{\boldsymbol{P}^p}{\boldsymbol{H}^p+\boldsymbol{P}^p} \tag{5}$$

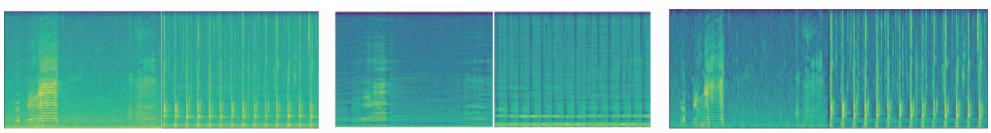
式中：$\boldsymbol{H}^p$ 为谐波增强部分的能量，$\boldsymbol{P}^p$ 为打击源增强部分的能量，上标 p 为能量系数。

最后，基于频谱并结合谐波增强部分和打击源部分的相对分量得到频谱谐波分量（spectrogram harmonic component，$\boldsymbol{C}_{\text{SH}}$）和频谱打击源分量（spectrogram percussive-source component，$\boldsymbol{C}_{\text{SP}}$），定义如下：

$$\boldsymbol{C}_{\text{SH}}=\boldsymbol{S}\otimes\boldsymbol{M}_{\text{H}} \tag{6}$$

$$\boldsymbol{C}_{\text{SP}}=\boldsymbol{S}\otimes\boldsymbol{M}_{\text{P}} \tag{7}$$

式中⊗表示两矩阵对应位置元素相乘，结果为同型矩阵。选取绵羊叫声和水滴声两个不同场景下，音频的频谱与其谐波分量和打击源分量进行对比，见图 1。



(a) 频谱　　(b) $\boldsymbol{C}_{\text{SH}}$　　(c) $\boldsymbol{C}_{\text{SP}}$

图 1　频谱分离前后特征对比

Fig. 1　Comparison of features before and after spectrogram separation

由图 1 可知：$\boldsymbol{C}_{\mathrm{SH}}$表现为平行于时间轴的水平线，$\boldsymbol{C}_{\mathrm{SP}}$表现为平行于频率轴的垂直线，且分离后得到的 $\boldsymbol{C}_{\mathrm{SH}}$和 $\boldsymbol{C}_{\mathrm{SP}}$特征中背景亮斑（通常由背景噪音产生）更少。

1.2 梅尔频谱的分离

频谱分离方法虽在一定程度上减少了特征图中的背景亮斑，但仍存在以下问题：1）频谱是赫兹尺度的特征矩阵，其包含的有效声学信息较少；2）不同音频样本中背景噪音的表现形式存在差异，频谱分离方法未能实现对背景噪音的针对性处理。有鉴于此，本文提出了一种梅尔频谱分离方法。

由于人耳的听觉范围与赫兹尺度的频率呈非线性关系，这导致了人耳对不同频率声音的感知灵敏度不同。基于这种特性，将声音频带从低到高按临界带宽由密到疏设置三角滤波器，即梅尔滤波器组[10]，其中心频率即为梅尔频率，梅尔频率 $\mathrm{Mel}(f)$ 与频率 f 间的转换关系为

$$\mathrm{Mel}(f)=\frac{1\ 000}{\log 2}\log\left(\frac{f}{1\ 000}+1\right) \tag{8}$$

首先，将频谱 $\boldsymbol{S}$ 通过上述梅尔滤波器组得到梅尔频谱 $\boldsymbol{S}_{\mathrm{m}}$，并将梅尔频谱的时域分量 $\boldsymbol{S}_{\mathrm{m}}(t)$ 和频域分量 $\boldsymbol{S}_{\mathrm{m}}(f)$ 的绝对值作为输入信号输入中值滤波器，其输出信号为梅尔频谱的谐波增强部分 $\boldsymbol{H}_{\mathrm{m}}$和打击源增强部分 $\boldsymbol{P}_{\mathrm{m}}$。

其次，将梅尔频谱的残差部分 $\boldsymbol{R}_{\mathrm{m}}$定义为

$$\boldsymbol{R}_{\mathrm{m}}=|\boldsymbol{S}_{\mathrm{m}}|-(\boldsymbol{P}_{\mathrm{m}}+\boldsymbol{H}_{\mathrm{m}}) \tag{9}$$

然后，基于梅尔频谱的谐波增强部分 $\boldsymbol{H}_{\mathrm{m}}$、打击源增强部分 $\boldsymbol{P}_{\mathrm{m}}$和残差部分 $\boldsymbol{R}_{\mathrm{m}}$，计算出其各自的相对分量 $\boldsymbol{M}_{\mathrm{Hm}}$、$\boldsymbol{M}_{\mathrm{Pm}}$和 $\boldsymbol{M}_{\mathrm{Rm}}$如下：

$$\boldsymbol{M}_{\mathrm{Hm}}=\frac{\boldsymbol{H}_{\mathrm{m}}^{p}}{\boldsymbol{H}_{\mathrm{m}}^{p}+\boldsymbol{P}_{\mathrm{m}}^{p}+\boldsymbol{R}_{\mathrm{m}}^{p}} \tag{10}$$

$$\boldsymbol{M}_{\mathrm{Pm}}=\frac{\boldsymbol{P}_{\mathrm{m}}^{p}}{\boldsymbol{H}_{\mathrm{m}}^{p}+\boldsymbol{P}_{\mathrm{m}}^{p}+\boldsymbol{R}_{\mathrm{m}}^{p}} \tag{11}$$

$$\boldsymbol{M}_{\mathrm{Rm}}=\frac{\boldsymbol{R}_{\mathrm{m}}^{p}}{\boldsymbol{H}_{\mathrm{m}}^{p}+\boldsymbol{P}_{\mathrm{m}}^{p}+\boldsymbol{R}_{\mathrm{m}}^{p}} \tag{12}$$

最后，利用梅尔频谱结合谐波增强部分、打击源增强部分和残差部分的相对分量得到梅尔频谱谐波分量 $\boldsymbol{C}_{\mathrm{MSH}}$（Mel-spectrogram harmonic component）、梅尔频谱打击源分量 $\boldsymbol{C}_{\mathrm{MSP}}$（Mel-spectrogram percussive-source component）和梅尔频谱残差分量 $\boldsymbol{C}_{\mathrm{MSR}}$（Mel-spectrogram residual component）分别为：

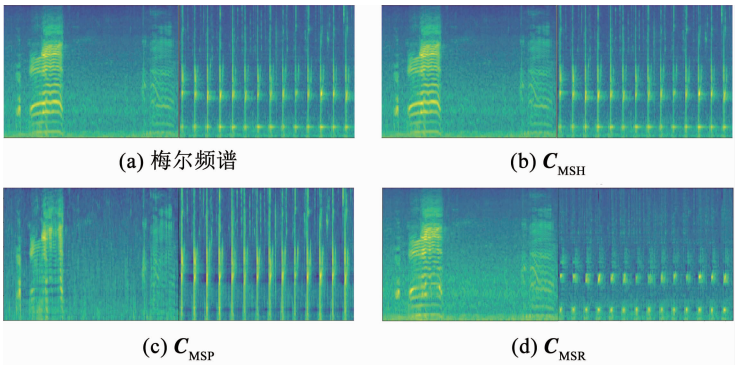
$$\boldsymbol{C}_{\mathrm{MSH}}=\boldsymbol{S}_{\mathrm{m}}\otimes\boldsymbol{M}_{\mathrm{Hm}} \tag{13}$$

$$\boldsymbol{C}_{\mathrm{MSP}}=\boldsymbol{S}_{\mathrm{m}}\otimes\boldsymbol{M}_{\mathrm{Pm}} \tag{14}$$

$$\boldsymbol{C}_{\mathrm{MSR}}=\boldsymbol{S}_{\mathrm{m}}\otimes\boldsymbol{M}_{\mathrm{Rm}} \tag{15}$$

为更直观地阐述梅尔频谱分离和频谱分离间的差别，同理选取上述 4 个不同场景下音频的梅尔频谱与其谐波分量、打击源分量和残差分量进行对比，见图 2。

对比图 1（a）和图 2（a）可知：1）相较于频谱，梅尔频谱中的亮斑更清晰，说明其中包含更多的声学信息；对比图 2 可知：2）梅尔频谱中的亮斑最清晰，但其包含有效特征信息和背景噪音信息；3）$\boldsymbol{C}_{\mathrm{MSH}}$中水平亮斑表现得更清晰，其虽增强了梅尔频谱中的谐波部分，但背景噪音的谐波部分同样也得到了增强；$\boldsymbol{C}_{\mathrm{MSP}}$中垂直亮斑表现得更清晰，其虽然增强了梅尔频谱中的打击源部分，但背景噪音的打击源部分同样也得到了增强；4）$\boldsymbol{C}_{\mathrm{MSR}}$中的亮斑较清晰，其不仅最大限度保留了有效特征信息，且针对性地减弱了背景噪音造成的亮斑；相比文献[3－5]中的特征融合或特征补偿方法，$\boldsymbol{C}_{\mathrm{MSR}}$针对性地抑制了背景噪音的影响。



(a) 梅尔频谱　(b) $\boldsymbol{C}_{\mathrm{MSH}}$　(c) $\boldsymbol{C}_{\mathrm{MSP}}$　(d) $\boldsymbol{C}_{\mathrm{MSR}}$

图 2 梅尔频谱分离前后特征对比

Fig. 2 Comparison of features before and after Mel-spectrogram separation

2 长距离自校正神经网络

近年来，利用卷积神经网络对声学信息进行提取已成为声学领域的重要研究方法之一[11－12]，必须指出的是：1）文献[6－8]仅采用单一的模型提取特征图中的有效信息，而减少了对特征图中浅层特征

的关注；2）声学特征图中的关键信息主要包含在频域中[13]，故增加对频域信息的关注，对特征图信息的提取有显著优势。

2.1　频域自校正算法

在图像识别领域，自校正结构常被应用为分组卷积的一个分支，该结构通过对图像的两个维度进行尺寸变换来间接扩大卷积核的感受野，故能更高效地提取图像中的上下文信息[14]。在声学场景分类领域，频谱常用来表示音频在时间和频率两个维度上的信息，且频域部分包含音频的大量关键信息，故提出一种频域自校正算法以期实现频域信息的高效采集，频域自校正算法的工作原理见图 3。

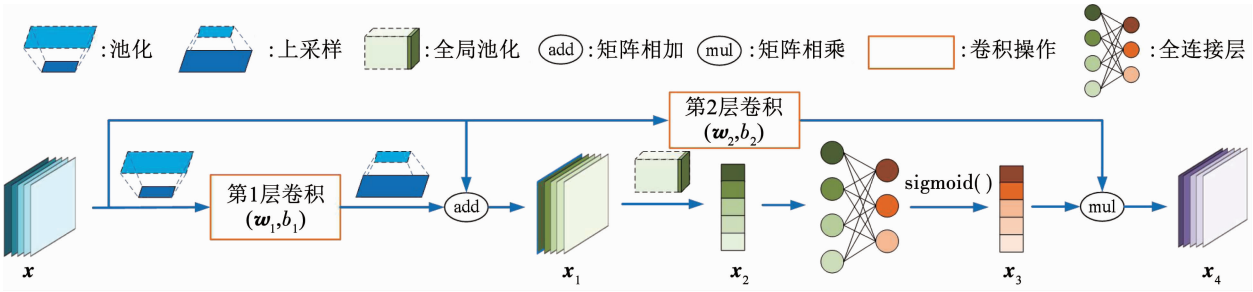


图 3　频域自校正结构示意

Fig. 3　Schematic diagram of frequency domain self-calibration structure

首先，通过池化运算将输入特征图 $\boldsymbol{x}$ 的频率维度压缩，再通过卷积层提取特征信息，从而间接扩大卷积核在频域上的感受野；其次，将输出结果通过上采样恢复成原始特征图的维度并与原始特征图求和得到 $\boldsymbol{x}_1$，以保留输入层特征图中的有效信息；然后，将结果通过全局池化压缩成一维向量 $\boldsymbol{x}_2$ 并通过全连接层和 sigmoid 函数将每个通道权值重新标定得到 $\boldsymbol{x}_3$；最后，将卷积后的原始特征图与权值向量 $\boldsymbol{x}_3$ 在通道维做乘积运算得到输出特征图 $\boldsymbol{x}_4$，以实现通道间的注意力增强。频域自校正结构的算法流程如下：

$$\boldsymbol{x}_1 = U\left[\boldsymbol{w}_1 D\left[\boldsymbol{x}\right]_{1\times r} + b_1\right]_{1\times r} + \boldsymbol{x} \tag{16}$$

$$\boldsymbol{x}_2 = F_{\mathrm{sq}}(\boldsymbol{x}_1) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{x}_1(i,j) \tag{17}$$

$$\boldsymbol{x}_3 = \sigma\left[\boldsymbol{w}_4 f(\boldsymbol{w}_3\boldsymbol{x}_2)\right] \tag{18}$$

$$\boldsymbol{x}_4 = (\boldsymbol{w}_2\boldsymbol{x} + b_2)\cdot\boldsymbol{x}_3 \tag{19}$$

式中：$D[\cdot]_{1\times r}$ 为池化运算，$U[\cdot]_{1\times r}$ 为上采样运算，其中 $1\times r$ 为池化核和上采样核的大小；$\boldsymbol{w}_1$ 和 b_1 为第 1 层卷积的权重矩阵和偏置值；$F_{\mathrm{sq}}(\cdot)$ 为挤压激励函数，其将尺寸为 $H\times W\times C$ 的特征图（H 为时域维度，W 为频域维度，C 为通道数）压缩为一维向量；$\sigma[\cdot]$ 为 sigmoid 函数；$f(\cdot)$ 为 ReLU 激活函数；$\boldsymbol{w}_3$ 和 $\boldsymbol{w}_4$ 分别为两个全连接层的权重矩阵；$\boldsymbol{w}_2$ 和 b_2 为第 2 层卷积的权重矩阵和偏置值。

基于上述分析，频域自校正算法通过压缩频率维度，提升了卷积核在频域的感受野，并通过残差结构和通道注意力增强机制，加强了深层特征与浅层特征间的关联度，故更有利于模型对有效信息的采集。

2.2　长距离增强结构

为进一步保留模型训练初期采集到的浅层特征信息并与之后采集到的深层特征信息互补偿，从而达到特征增强的作用，提出了一种长距离增强结构，下面介绍其具体工作原理。

首先，设置 n 组卷积组用于采集特征图中的浅层特征信息（n 为大于 1 的整数）；其次，基于频域自校正结构设置 n 组频域自校正模块（frequency-domain self-calibration block，FSC Block）用于采集特征图中的深层特征信息；最后，利用残差操作将第 $n-1$ 个卷积组的输出特征与第 1 个 FSC Block 的输出特征进行叠加融合，第 $n-2$ 个卷积组的输出特征与第 2 个 FSC Block 的输出特征进行叠加融合，以此类推，直到第 1 个卷积组的输出特征与第 $n-1$ 个 FSC Block 的输出特征进行叠加融合，以实现网络模型的长距离增强，期间分别采用卷积运算和上采样运算使二者的通道数及分辨率保持一致。当 $n=3$ 时，长距离增强结构示意见图 4。

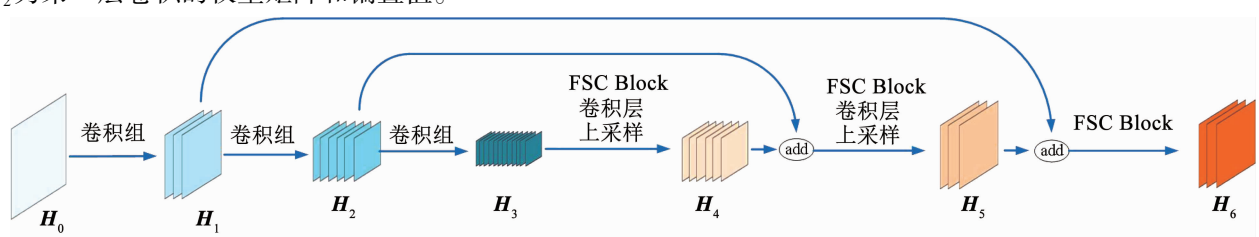


图 4　长距离增强结构示意

Fig. 4　Schematic diagram of long-distance reinforcement structure

图 4 中，$\boldsymbol{H}_0$ 为输入层，$\boldsymbol{H}_1$、$\boldsymbol{H}_2$ 和 $\boldsymbol{H}_3$ 为经过卷积组操作的输出层，$\boldsymbol{H}_4$、$\boldsymbol{H}_5$ 为经过 FSC Block、卷积层和上采样操作的输出层，$\boldsymbol{H}_6$ 为经过 FSC Block 的输出层。卷积组中均包含两个卷积层和一个 2 × 2 池化层。当 n 设置为不同值时，长距离增强结构的各项参数设置见表 1。

表 1 长距离增强结构参数设置

Tab. 1 Parameter setting of long-distance reinforcement structure

卷积组数	FSC Block 个数	残差操作数/个	残差操作
2	2	1	$\boldsymbol{H}_1+\boldsymbol{H}_3$
3	3	2	$\boldsymbol{H}_1+\boldsymbol{H}_5,\boldsymbol{H}_2+\boldsymbol{H}_4$
4	4	3	$\boldsymbol{H}_1+\boldsymbol{H}_7,\boldsymbol{H}_2+\boldsymbol{H}_6,\boldsymbol{H}_3+\boldsymbol{H}_5$
⋮	⋮	⋮	⋮
n	n	$n-1$	$\boldsymbol{H}_1+\boldsymbol{H}_{2n-1},\boldsymbol{H}_2+\boldsymbol{H}_{2n-2},\cdots,\boldsymbol{H}_{n-1}+\boldsymbol{H}_{n+1}$

由表 1 可知，当设置 n 组卷积组时，第 i 个残差操作的输出层 $\boldsymbol{H}_i^{\mathrm{Res}}$ 定义为

$$\boldsymbol{H}_i^{\mathrm{Res}}=\boldsymbol{H}_i+\boldsymbol{H}_{2n-i},\ i=1,2,\cdots,n-1 \tag{20}$$

由于 FSC Block 操作输出的特征图具有不同的分辨率，且不同分辨率的特征图中包含的特征信息均存在一定的差异。为实现不同分辨率特征信息的互补偿自适应，并提高对小目标的检测精度，利用多尺度特征融合方法[15]，将 FSC Block 操作的输出层尺寸通过池化操作进行统一，并将结果在通道维度进行合并。

综上所述，长距离增强结构将模型采集到的浅层信息和深层信息通过叠加融合的方式进行互补偿，并利用多尺度特征融合方法将不同分辨率的特征进行通道合并，以实现不同分辨率信息之间的互补偿。

3 实 验

3.1 实验数据集

为进一步验证基于梅尔频谱分离和 LSCNet 的声学场景分类方法的有效性，利用城市声音标准数据集 Urbansound8K[16] 和环境音频数据集 ESC-50[17] 开展声学场景分类研究。其中，Urbansound8K 数据集包含 8 732 个长度不同的音频样本，共分为 10 个类别。ESC-50 数据集包含 2 000 个长度均为 5 s 的音频样本，共分为 50 个类别（每个类别包含 40 个样本）。

3.2 网络框架与参数配置

实验中根据特征图的尺寸分别搭建了 3 组参数不同的网络模型，网络模型参数见表 2。参数配置为：训练时，设置批处理尺寸（batch-size）为 40 时模型分类效果最优，采用 Adam 为优化器，设置初始学习率为 0.01，且训练轮次（epoch）设置为 300 时，模型的收敛效果最佳。

表 2 网络模型参数

Tab. 2 Parameters of network models

网络模型	卷积层数/个	卷积组数/个	残差操作数/个	参数量	模型大小/MB
LSCNet-13	13	2	1	540 516	6.57
LSCNet-19	19	3	2	883 026	10.68
LSCNet-25	25	4	3	1 353 120	16.29

3.3 实验结果分析

3.3.1 模型深度实验

模型训练的过程中，模型深度会对其提取特征的能力产生影响，从而在一定程度上影响模型的分类准确率。为探究最佳网络深度，并比较 LSCNet 与已有自校正网络（self-calibration network，SCNet）的分类性能，构建了表 2 中 LSCNet-13、LSCNet-19 和 LSCNet-25 网络模型，设置了相同层数的 SCNet-13、SCNet-19 和 SCNet-25 模型作为对照组；采用梅尔频谱特征并基于 Urbansound8K 数据集开展声学场景分类实验，具体实验结果见表 3。

表 3 不同深度模型的分类准确率对比

Tab. 3 Classification accuracy comparison of different deep models %

深度模型	分类准确率	深度模型	分类准确率
SCNet-13	85.5	LSCNet-13	87.2
SCNet-19	86.8	LSCNet-19	88.6
SCNet-25	86.3	LSCNet-25	87.6

由表 3 可知：1）相较于相同层数的 SCNet 模型，LSCNet-13、LSCNet-19 和 LSCNet-25 模型的分类准确率分别提高了 1.7%、1.8% 和 1.3%；2）当设置 19 层卷积层时，LSCNet 模型取得最高的分类准确率为 88.6%，故当设置 3 组卷积组时，模型的分类性能最优。

3.3.2 特征分离对比实验

为验证梅尔频谱分离方法相较于频谱分离方法能针对性地减少背景噪音的影响并提高模型的分类准确率，采用 7 种特征作为输入特征图，并采用 LSCNet-19 作为训练模型，分别基于 Urbansound8K 数据集和 ESC-50 数据集开展实验，实验结果见表 4。

表 4　不同特征的分类准确率对比

Tab. 4　Classification accuracy comparison of different features　%

数据集	频谱	C_{SH}	C_{SP}	梅尔频谱	C_{MSH}	C_{MSP}	C_{MSR}
Urbansound8K	85.7	83.3	84.3	88.6	89.4	89.0	90.1
ESC-50	82.2	76.8	79.9	83.3	86.0	85.0	88.0

由表 4、图 5、6 可得出以下结论：1）相较于频谱，利用梅尔频谱输入 LSCNet-19 模型进行训练时，模型的分类准确率在 Urbansound8K 和 ESC-50 数据集上分别提高了 2.9% 和 1.1%；2）相较于频谱，将频谱分离后得到的 C_{SH} 和 C_{SP} 输入模型进行训练时，模型的分类准确率在 Urbansound8K 数据集上分别降低了 2.4% 和 1.4%，在 ESC-50 数据集上分别降低了 5.4% 和 2.3%；3）相较于梅尔频谱，将梅尔频谱分离后得到的 C_{MSH}、C_{MSP} 和 C_{MSR} 输入模型进行训练时，模型的分类准确率在 Urbansound8K 数据集上分别提高了 0.8%、1.4% 和 1.5%，在 ESC-50 数据集上分别提高了 2.7%、1.7% 和 4.7%；4）相较于频谱分离后得到的 C_{SH} 和 C_{SP}，将梅尔频谱分离后得到的 C_{MSH}、C_{MSP}、和 C_{MSR} 输入模型进行训练时，模型的分类准确率在两个数据集上平均提高了 5.7% 和 7.9%；5）相较于其他声学特征，C_{MSR} 在含背景噪音较多的空调、儿童玩耍和街头音乐等多个类别的子分类准确率达到最高，进一步验证了其可以针对性地减少背景噪音的影响。

上述结论也进一步验证了梅尔频谱分离方法相较于频谱分离方法得到的特征具有良好的分类性能，且 C_{MSR} 特征能够针对性地减少背景噪音对模型分类性能的影响，具有最优的分类性能。

3.3.3　分类准确率对比实验

为了验证基于梅尔频谱分离和 LSCNet 的声学场景分类方法的有效性，基于 Urbansound8K 和 ESC-50 数据集开展了声学场景分类实验，并与已有声学场景分类模型进行比较，具体结果见表 5。

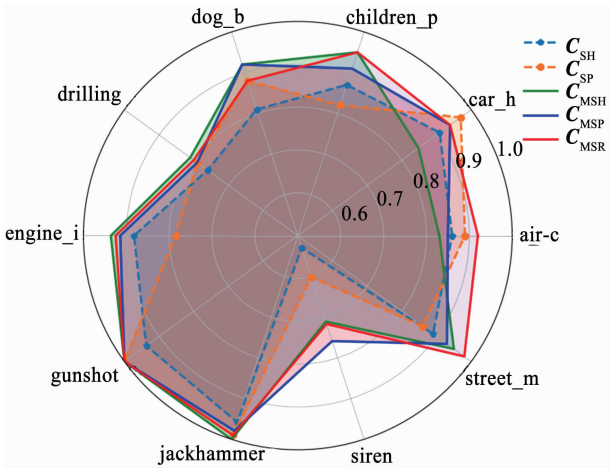


图 5　不同特征的子分类准确率雷达图

Fig. 5　Sub-classification accuracy radar chart of different features

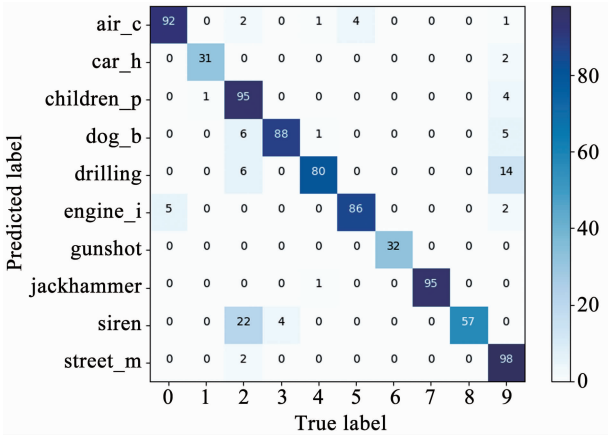


图 6　C_{MSR} 准确率为 90.1% 的混淆矩阵

Fig. 6　Confusion matrix of C_{MSR} with 90.1% accuracy

表 5　不同声学场景分类模型的准确率对比

Tab. 5　Accuracy comparison of different acoustic scene classification models

文献	网络模型	声学特征	模型参数量	Urbansound8K/%	ESC-50/%
文献[16]（基线系统）	Random Forest	MFCC			44.3
文献[6]	CNN + Mixup	Mels + Gts	973 578	83.7	83.9
文献[7]	DCNN + Attention	LM + Gts	1 497 290		86.5
文献[8]	CNN	MFCC-LM-CST	1 243 882	89.3	85.6
本文模型	LSCNet-19	C_{MSR}	883 026	90.1	88.0

由表 5 可得出以下结论：1）利用 C_{MSR} 特征输入 LSCNet-19 网络模型，模型的分类准确率在两个数据集下分别达到了 90.1% 和 88.0%；2）相较于基线系统，本文提出的 LSCNet-19 模型在 C_{MSR} 特征下，基于 ESC-50 数据集的分类准确率提高了 35.7%；3）相较于文献[6－8]提出的系统，基于 ESC-50 数据集，本文提出的 LSCNet-19 模型在 C_{MSR} 特征下分类准确率分别提高了 4.1%、1.5% 和 2.4%；4）相较于文

献[6,8]提出的系统,基于 Urbansound8K 数据集,本文提出的 LSCNet-19 模型在 $\boldsymbol{C}_{\mathrm{MSR}}$ 特征下分类准确率分别提高了 6.4% 和 0.8%;5)相较于文献[6-8]提出的系统,本文模型包含的参数量较少,即模型的复杂度较低。

综合实验结果表明:1)相较于频谱分离方法,梅尔频谱分离方法分离出的 $\boldsymbol{C}_{\mathrm{MSR}}$ 特征能够针对性地减少背景噪音的影响,从而提高模型的分类准确率;2)本文提出的 LSCNet 既实现了对特征图中频域信息的关注,又将模型训练中采集到的浅层特征信息和深层特征信息相互补偿,使模型具有良好的分类性能。

4 结 论

通过对 Urbansound8K 和 ESC-50 数据集开展声学场景分类实验与分析,得出主要结论:

1)相较于频谱分离后得到的 $\boldsymbol{C}_{\mathrm{SH}}$ 和 $\boldsymbol{C}_{\mathrm{SP}}$,将梅尔频谱分离后得到的 $\boldsymbol{C}_{\mathrm{MSH}}$、$\boldsymbol{C}_{\mathrm{MSP}}$、和 $\boldsymbol{C}_{\mathrm{MSR}}$ 输入模型进行训练时,模型的分类准确率在两个数据集上平均提高了 5.7% 和 7.9%;且 $\boldsymbol{C}_{\mathrm{MSR}}$ 在含背景噪音较多的空调、儿童玩耍和街头音乐等多个类别的子分类准确率达到最高,验证了其可以针对性地减少背景噪音的影响。

2)相较于自校正网络,LSCNet 实现了对特征图中频域信息的关注,在 Urbansound8K 数据集下,模型准确率平均提升了 1.6%,且当模型层数设置为 19 时分类效果达到最佳,LSCNet-19 在两个数据集下的准确率分别达到 90.1% 和 88%,进一步验证了该方法的有效性。

3)相较于文献[6-8]提出的系统,基于 ESC-50 数据集,本文提出的 LSCNet-19 模型在 $\boldsymbol{C}_{\mathrm{MSR}}$ 特征下分类准确率分别提高了 4.1%、1.5% 和 2.4%,相较于文献[6,8]提出的系统,基于 Urbansound8K 数据集,本文提出的 LSCNet-19 模型在 $\boldsymbol{C}_{\mathrm{MSR}}$ 特征下分类准确率分别提高了 6.4% 和 0.8%,且模型的复杂度更低。

参考文献

[1]易江燕, 陶建华, 刘斌, 等. 基于迁移学习的噪音鲁棒语音识别声学建模[J]. 清华大学学报, 2018, 58(1): 56
YI Jiangyan, TAO Jianhua, LIU Bin, et al. Transfer learning for acoustic modeling of noise robust speech recognition[J]. Journal of Tsinghua University, 2018, 58(1): 56. DOI: 10.16511/j.cnki.qhdxxb.2018.21.001

[2]刘亚荣, 黄昕哲, 谢晓兰, 等. 美尔谱系数与卷积神经网络相组合的环境声音识别方法[J]. 信号处理, 2020, 36(6): 1022
LIU Yarong, HUANG Xinzhe, XIE Xiaolan, et al. Environmental sound recognition method combining Meir spectral coefficients and convolutional neural network[J]. Journal of Signal Processing, 2020, 36(6): 1022. DOI: 10.16798/j.issn.1003-0530.2020.06.025

[3]赵薇, 靳聪, 涂中文, 等. 基于多特征融合的 SVM 声学场景分类算法研究[J]. 北京理工大学学报, 2020, 40(1): 70
ZHAO Wei, JIN Cong, TU Zhongwen, et al. Support vector machine for acoustic scene classification algorithm research based on multi-features fusion[J]. Transactions of Beijing Institute of Technology, 2020, 40(1): 70. DOI:10.15918/j.tbit1001-0645.2018.171

[4]张科, 苏雨, 王靖宇, 等. 基于融合特征以及卷积神经网络的环境声音分类系统研究[J]. 西北工业大学学报, 2020, 38(1): 164
ZHANG Ke, SU Yu, WANG Jingyu, et al. Environment sound classification system based on hybrid feature and convolutional neural network[J]. Journal of Northwestern Polytechnical University, 2020, 38(1): 164. DOI: 10.1051/jnw-pu/20203810162

[5]CAO G, AKBAS C E, CETIN A E. Recognition of vessel acoustic signatures using non-linear teager energy based features[C]//2016 International Workshop on Computational Intelligence for Multimedia Understanding. Reggio Calabria: IEEE, 2016:2. DOI: 10.1109/IWCIM.2016.7801190

[6]ZHANG Zhichao, XU Shugong, CAO Shan, et al. Deep convolutional neural network with mixup for environmental sound classification[C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Guangzhou: Springer, 2018: 362. DOI: 10.1007/978-3-030-03335-431

[7]ZHANG Zhichao, XU Shugong, ZHANG Shunqing, et al. Learning attentive representations for environmental sound classification[J]. IEEE Access, 2019(7): 130337. DOI: 10.1109/ACCESS.2019.2939495

[8]SU Yu, ZHANG Ke, WANG Jingyu, et al. Performance analysis of multiple aggregated acoustic features for environment sound classification[J]. Applied Acoustics, 2020, 158: 107050. DOI: 10.1016/j.apacoust.2019.107050.

[9]FITZGERALD D. Harmonic/Percussive separation using median filtering[C]//13th International Conference on Digital Audio Effects (DAFX10). Graz: [s.n.], 2010

[10]VIRTANEN T, PLUMBLEY M D, ELLIS D. Computational analysis of sound scenes and events[M]. [S.l.]: Springer, 2018: 77

[11]程艳芬, 陈垚鑫, 陈逸灵, 等. 嵌入注意力机制并结合层级上下文的语音情感识别[J]. 哈尔滨工业大学学报, 2019, 51(11): 100
CHENG Yanfen, CHEN Yaoxin, CHEN Yiling, et al. Speech emotion recognition with embedded attention mechanism and hierarchical context[J]. Journal of Harbin Institute of Technology, 2019, 51(11): 100. DOI: 10.11918/j.issn.0367-6234.201905193

[12]HUANG Zilong, LIU Chen, FEI Hongbo, et al. Urbansound classification based on 2-order dense convolutional network using dual features[J]. Applied Acoustics, 2020, 164:107243. DOI: 10.1016/j.apacoust.2020.107243

[13]PHAYE S S R, BETENYOS E, WANG Ye. SubSpectralNet-using sub-spectrogram based convolutional neural networks for acoustic scene classification[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 828. DOI: 10.1109/ICASSP.2019.8683288

(下转第 123 页)