引用格式：姚琨, 杨吉斌, 张雄伟, 等. 基于多分辨率时频特征融合的声学场景分类[J]. 声学技术, 2020, **39**(4): 494-500. [YAO Kun, YANG Jibin, ZHANG Xiongwei, et al. Acoustic scene classification based on multi-resolution time-frequency feature fusion[J]. Technical Acoustics, **39**(4): 494-500.]
DOI: 10.16300/j.cnki.1000-3630.2020.04.019

基于多分辨率时频特征融合的声学场景分类

姚 琨，杨吉斌，张雄伟，郑昌艳，孙 蒙
(陆军工程大学，江苏 南京 210007)

摘要：声学场景分类是计算机听觉中最难的任务之一，在单一特征条件下采用基本的卷积神经网络相对于传统的分类方法精度已经有所提升，但是效果依然不够理想。针对这一问题，在卷积神经网络框架下，提出了一种基于时频特征融合的声学场景分类方案。在分类模型构建方面，提出一种多分辨率卷积池化方案，构造多分辨率卷积神经网络，以更好地适应提取特征的时频结构；在特征选取方面，融合低层次包络特征对数——Mel 子带能量和高层次结构特征——非负矩阵分解系数矩阵，把两种二维特征堆叠为三维特征送入分类模型。在 2017 年和 2018 年声学场景分类和事件检测挑战赛的开发数据集上进行了训练和测试。实验结果表明，文中提出方案比基线系统的分类精度分别提高 7.5%和 10.3%，可有效改善分类效果。

关键词：声学场景分类；多分辨率卷积神经网络；时频特征融合；时频结构；非负矩阵分解
中图分类号：TN911.7 **文献标识码：**A **文章编号：**1000-3630(2020)-04-0494-07

Acoustic scene classification based on multi-resolution time-frequency feature fusion

YAO Kun, YANG Jibin, ZHANG Xiongwei, ZHENG Changyan, SUN Meng
(*Army Engineering University*, *Nanjing* 210007, *Jiangsu*, *China*)

Abstract: Acoustic scene classification is one of the most difficult tasks in computer hearing. It is difficult to achieve good classification performance by using basic convolutional neural network structure under the condition of single feature. To solve this problem, this paper proposes an acoustic scene classification scheme based on time-frequency feature fusion and multi-resolution convolutional neural network. In the model design, a multi-resolution pooling scheme is adopted to construct a multi-resolution convolutional neural network, which can better adapt to the time-frequency structure of feature extraction. In the feature extraction, the Log Mel-band energies of low level envelope features and the non-negative matrix decomposition coefficient matrix of high level structure features are fused into three dimensional features to input the classification model. Training and testing are carried out on the development data sets of the acoustic scene classification and event detection challenge in 2017 and 2018. The experimental results show that the classification accuracy of the proposed scheme is 7.5% and 10.3% higher than the classification accuracy of the baseline system respectively.

Key words: acoustic scene classification; multi-resolution convolutional neural network; time-frequency feature fusion; time-frequency structure; non-negative matrix factorization

0 引 言

智能系统要想充分利用声音信息，不仅要识别语音[1]和音乐[2]等单一的声音信号，还要充分利用周围环境中的声音信息，声学场景分类(Acoustic Scenes Classification, ASC)的目的就是利用周围环境中的声音信息对预先定义的有限场景类型进行分类。近几年，ASC 已经成为声学场景和事件的分类和检测(Detection and Classification of Acoustic Scenes and Events, DCASE)挑战赛中最热门的任务之一[3-6]，在智能穿戴设备的自动模式调整等方面有着广泛的应用前景。

ASC 的目的是通过选择一个语义标签来描述音频流表达的声学环境，是机器学习中典型的单标签分类任务。与一般的音频分类任务类似，ASC 任务的关键就在于模型设计和音频信号特征选取两个方面。

在特征选取方面，多采用音频信号的二维时频表示，时间轴以帧为单位构造长帧，频率轴为帧信号短时傅里叶变换的降维表示。其中最为典型的有

收稿日期：2019-03-03; **修回日期：**2019-04-30
基金项目：国家自然科学基金(61471394)、江苏省优秀青年基金(BK20180080)资助项目。
作者简介：姚琨(1994－), 男, 山西长治人, 硕士研究生, 研究方向为智能信息处理。
通讯作者：杨吉斌, E-mail: yjbice@sina.com

梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)[7]、对数梅尔子带能量(Log Mel-band energies, LME)等特征。它们都是对音频信号谱包络的整体描述，同时经梅尔滤波器处理过后符合人的听觉感受，具有较好的区分度和鲁棒性。

文献[8]对场景音频信号的声谱图进行非负矩阵分解(Non-negative Matrix Factorization, NMF)，输出一组激活函数。这组激活函数对基函数的贡献进行编码，从而对声学场景的属性进行建模。

在分类模型方面，由于 ASC 是语音识别和音乐识别的拓展研究，在这些任务中广泛使用的模型[9-10]也被应用到 ASC 任务中，如隐马尔可夫模型(Hidden Markov Model, HMM)、高斯混合模型(Gaussian Mixed Model, GMM)和支持向量机(Support Vector Machine, SVM)。近年来，由于深度神经网络对非线性问题强大的表征能力，在语音和图像方面取得巨大的成功，人们开始把深度学习的方法应用到了 ASC 任务中来，包括深度神经网络(Deep Neural Network, DNN)、卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)等。

Eghbal-Zadeh 等[11]提出了两种方案并且包揽 DCASE 2016 中 ASC 任务的前两名。第一种方案是用音频信号左声道、右声道、两声道的均值和差值 4 种 MFCC 特征分别训练身份认证矢量(Identity Vector, i-vector)，最后以分类得分的均值作为结果输出。第二种方案是在第一种方案的基础上，融合以声谱图为特征的 CNN 网络。Zheng 等[12]在对音频信号进行短时傅里叶变换时，通过改变滑动窗的宽度、FFT 的点数、窗的重叠量、降维的尺度等提取不同的声谱图特征，并融合常数 Q 变换(Constant Q Transform, CQT)特征训练 CNN 模型。Han 等[13]以 LME 为基础，从左右声道、左右声道和或差的音频中提取 LME 及单声道谐波和冲击响应特征等，分别训练视觉几何组(Visual Geometry Group, VGG)结构的模型，最后取所有模型分类精度的均值作为最终输出。

虽然以上方案都取得了较好的分类效果，但还存在以下几个问题：

(1) 声谱图特征的维度过高，需要以更深层的网络以及更大的数据量作为支撑，极大地耗费计算资源，且含有大量的冗余信息，类别间的区分度也不够高。

(2) 单一的低维度特征对音频信号的表征能力有限，现有的特征融合方法可解释性较差。

(3) 分类模型不具针对性，将在图像任务中表现较好的模型(如 VGG 等)进行简单的移植，并不能考虑到 ASC 任务中输入声音特征的时频结构。

为解决以上问题，本文以 CNN 为基本识别模型，在时频特征基础上进行了包络/结构特征融合、适应时频结构特征的多分辨率模型设计等融合处理。模型方面，本文基于 CNN 模型提出一种适应声音时频结构的多分辨率卷积池化方案，并构造多分辨率卷积神经网络(Multi-resolution Convolutional Neural Network, MCNN)模型；特征选择方面，采用低级包络特征 LME 辅以高级结构特征 NMF，在特征表征能力增强的同时，数据量及冗余信息增加较少，从而在模型和特征两方面共同完成对时频特征的融合。在 DCASE2017 和 DCASE2018 数据集上进行实验，结果验证了本文提出方案的有效性。

1　方案框架

本文采用特征提取-模型分类的实现框架。声音特征提取和模型分类系统的流程框图如图 1 所示。特征提取部分分为两个分支，分别提取 LME 特征和 NMF 特征。提取 NMF 特征时，首先提取音频样本的声谱图特征，然后利用由所有训练集样本训练得到的字典矩阵 $\boldsymbol{W}$，对提取到的声谱图特征进行 NMF，得到系数矩阵 $\boldsymbol{H}$ 作为 NMF 特征。特征融合部分，借鉴图像任务中 RGB 三通道的方法，把二维的 LME 和 NMF 特征堆叠为三维特征。最后把得到的三维特征送入 MCNN 模型得到分类结果。实

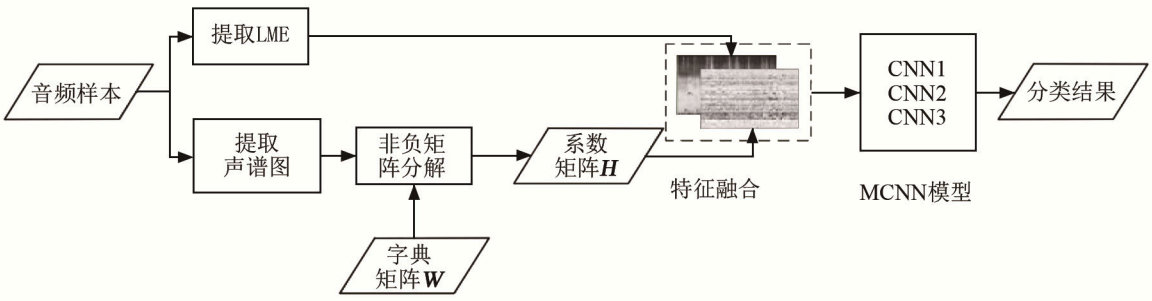


图 1　声音特征提取和模型分类系统的流程框图

Fig.1　Block diagram of voice feature extraction and model classification system

验证明，对于这两类特征，本文的融合方法要优于简单拼接和模型结果融合的方法。

2 MCNN 模型

2.1 适应声音时频结构的 CNN 框架

对声音信号进行短时处理后，其前后相邻帧的频谱分布高度相关，而频率轴上在经过后续提取手工特征时的降维处理后，彼此相对独立。针对这一特点，基于 CNN 框架，提出一种适应声音时频结构的池化策略，即池化操作时只沿着时间轴进行(频率轴采样因子为 1)，并且在经过两次池化后，特征在时间轴上维度变为 1 而频率轴维度保持不变。依此方法，可以设计出适应声音时频结构的 CNN 模型。

图 2 为本文模型第二层卷积的卷积核在输入特征图上的感受野示意图，深色部分为实际特征图，白色为“补 0”部分。由图 2 可知，CNN 提取到的高级特征是由输入特征图感受野上的数据经过卷积层加权求和及池化层采样得到的，卷积核在输入特征图上感受野的范围取决于卷积核和池化窗的大小及形状。采用带状的感受野，可以更好地适应输入特征图的时频结构。

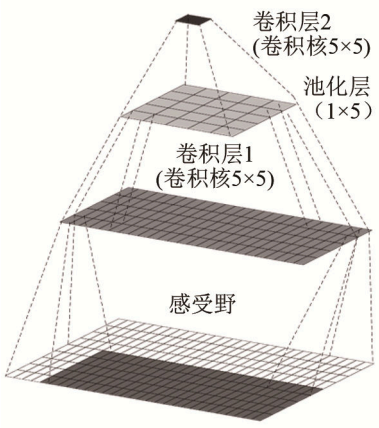


图 2 时频感受野示意图

Fig.2 Schematic diagram of time-frequency sensing field

2.2 MCNN

改变卷积核和池化窗的大小会影响 CNN 提取到的高级特征，从而直接影响到分类结果。采用不同大小的卷积核和池化窗可以捕捉到更全面的信息，小的卷积核和池化窗对应输入特征小的感受野，从而可以捕捉到特征图上的细节信息，大的卷积核和池化窗则相反，二者互为补充。

为此，我们在 CNN 中采用不同尺寸的卷积核和池化窗组合，力求同时捕捉细节和宏观信息。由于第二层卷积输出的高级特征来自于不同的子模型，每个子模型输出特征由同一输入特征图上不同大小的感受野提取得到，所以称之为多分辨率卷积神经网络(MCNN)，模型结构和参数如表 1 所示。其中，在输入层分别采用了 5×5 卷积+1×5 池化、7×7 卷积+1×10 池化、3×3 卷积+1×2 池化操作，在中间层分别采用了 5×5 卷积+1×100 池化、7×7 卷积+1×50 池化、3×3 卷积+1×250 池化。这些参数都是经过组合实验优选出来的结果。

表 1 MCNN 模型结构

Table 1 MCNN model structure

Input data (40×500×N)		
5×5Conv2D-32-BN- Relu	7×7Conv2D-32-BN-Relu	3×3Conv2D-32-BN-Relu
1×5 MaxPooling-Dropout(0.3)	1×10 MaxPooling-Dropout(0.3)	1×2 MaxPooling-Dropout(0.3)
5×5Conv2D-64-BN- Relu	7×7Con2D-64-BN-Relu	3×3Conv2D-64-BN-Relu
1×100 MaxPooling-Dropout(0.3)	1×50 MaxPooling-Dropout(0.3)	1×250 MaxPooling-Dropout(0.3)
Flatten	Flatten	Flatten
Concatenate		
Dense(100) -Relu- Dropout(0.5)		
Dense(10)- Softmax		
Output scene		

3 特征选取

LME 特征能反映声谱图帧内各频带的局部统计特性，对 ASC 任务中的音频样本有较好的低维表征能力，在近年来的 ASC 任务中使用最为广泛[13-14]。因此本文对特征的研究以 LME 为基础，针对 LME 特征的不足，提出 LME 与 NMF 特征的融合思路。

3.1 LME 特征

把音频信号分帧、加窗后，对每帧信号进行离散傅里叶变换，可得：

$$S_i(k)=\sum_{n=1}^{N} s_i(n)h(n)\mathrm{e}^{-\mathrm{j}2\pi kn/N} \qquad 1\leqslant k\leqslant N \tag{1}$$

其中：$s_i(n)$ 为第 i 帧时域信号，$h(n)$ 为长度为 N 的短时窗系数，N 是离散傅里叶变换的长度。

Mel 滤波器组的频率响应为

$$H_m(k)=\begin{cases}0, & k<f(m-1)\\ \dfrac{2(k-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m-1)\leqslant k<f(m)\\ \dfrac{2(f(m+1)-k)}{(f(m+1)-f(m-1))(f(m)-f(m-1))}, & f(m)\leqslant k<f(m+1)\\ 0, & f(m+1)\leqslant k\end{cases} \tag{2}$$

且满足 $\sum_{m=0}^{M-1} H_m(k)$，M 为滤波器个数。

计算每个滤波器输出的对数能量：

$$s(m)=\ln\left[\sum_{k=0}^{N-1}|S_i(k)|^2 H_m(k)\right],\ 0\leqslant m\leqslant M \quad (3)$$

最后构成的特征图横轴为时间轴，纵轴为各个滤波器下输出的对数能量，维度等于滤波器个数。

3.2　NMF 特征

NMF 是一种无监督的矩阵分解方法，它使分解后的所有分量均为非负值，并且同时实现非线性的维数约减。给定一个非负矩阵 $\boldsymbol{V}$，它的目标就是分解得到一个非负的字典矩阵 $\boldsymbol{W}$ 和一个系数矩阵 $\boldsymbol{H}$，即求解如下优化问题：

$$\min D_\beta(\boldsymbol{V}|\boldsymbol{WH}) \quad st.\ \ \boldsymbol{W},\boldsymbol{H}\geqslant 0 \quad (4)$$

其中，D_β 表示分离方式。本文中，我们把 β 设为 2，则 D_β 为欧氏距离。利用 NMF 得到的系数矩阵，作为 NMF 特征使用。

3.3　融合特征

LME 特征每一维都对应相应 Mel 滤波器输出的对数能量，在较好地表征音频信号的同时，降低了对应声谱图的维数，去除了大量冗余信息，减轻了计算复杂度。但是，使用 Mel 滤波器对声谱图进行平滑处理使得谐波结构(如图 3 中的机场音频样本声谱图的低频部分)这样的细节不再显著。图 4 为该样本经过 40 维 Mel 滤波器组后的输出图像。从图 4 可以看出，原本低频部分存在的明显的谐波结构已经完全无法观察到。实际上，Mel 滤波器组对其相应频带内频谱幅度进行了加权求和，从而实现降维处理，这就导致不同的频谱可能对应相同的滤波器输出，造成不同类别音频信号的 LME 特征在某些子区域上很可能相似或相同。

有效弥补 LME 特征中丢失的细节，可以将声谱图上的典型谐波结构作为细节信息补充上去，以增加原始 LME 特征的区分度。为此，提出一种为

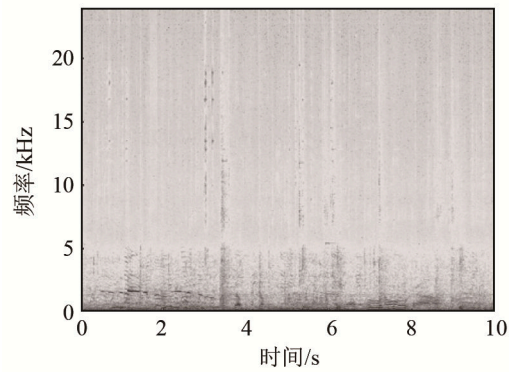


图 3　机场音频样本声谱图

Fig.3　Spectrogram of airport audio sample

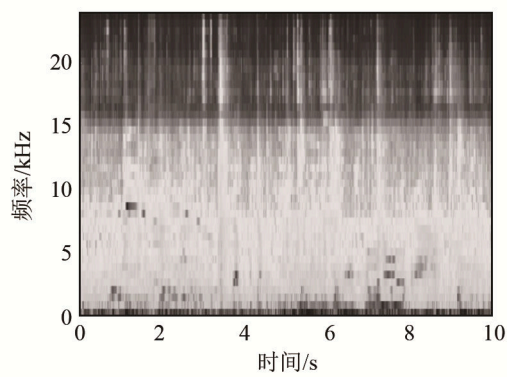


图 4　LME 特征图

Fig.4　LME feature map

细节特征与整体效果特征相融合的办法，对声谱图进行非负矩阵分解，之后将系数矩阵作为细节特征和 LME 特征相融合，在较少增加数据量和冗余信息的同时使得提取的特征具有更强的鲁棒性。

NMF 特征采用乘性更新规则来优化提取。特征提取过程如下：

(1) 把声谱图矩阵作为待分解矩阵(1 025×500)，将所有训练集样本的声谱图扩展成 1 025×500N 的二维矩阵，其中 N 为训练集的样本数。采用 NMF，提取时间轴上的基，分解得到大小为 1 025×40 的字典矩阵 $\boldsymbol{W}$，如图 5 所示。图 5 横轴上的每一维(时间轴上的每一维)对应为字典的一个基。

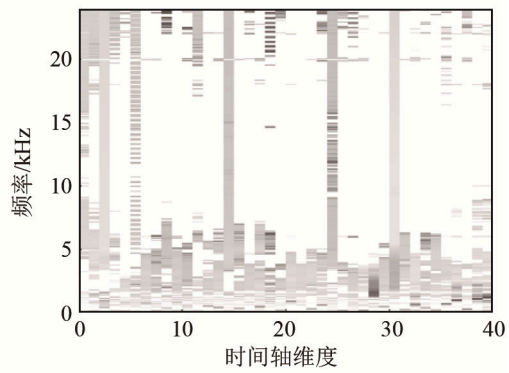


图 5　字典矩阵 $\boldsymbol{W}$

Fig.5　Dictionary matrix $\boldsymbol{W}$

(2) 固定由步骤(1)得到的字典矩阵 $\boldsymbol{W}$，对每个样本的声谱图矩阵作 NMF，得到 40×500 的非负矩阵系数 $\boldsymbol{H}$，如图 6 所示。这个矩阵 $\boldsymbol{H}$ 作为样本特征的一部分使用。

NMF 分解得到的系数矩阵 $\boldsymbol{H}$ 保持了声谱的低频细节结构。图 7 为原始声谱图和 NMF 重构的声谱图。对比图 7(a)、7(b)可以看出，重构声谱图在高频部分存在一定的失真，而低频部分的谱细节还原较好。从矩形框中的部分可以看出，声谱图低频段的谐波结构基本没有失真。因此，利用系数矩阵

$\boldsymbol{H}$，可以构造表征细节的低维特征。

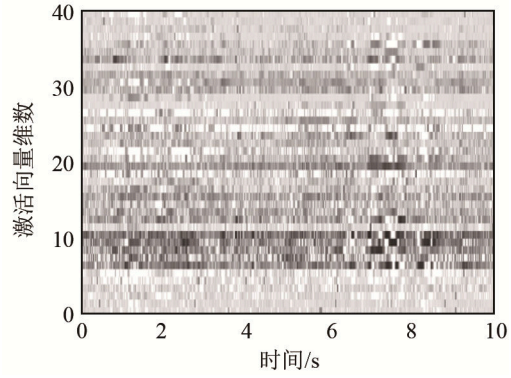
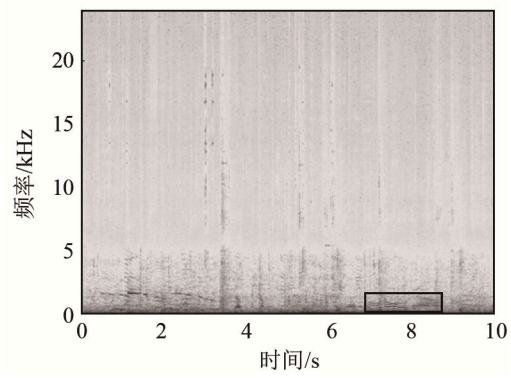
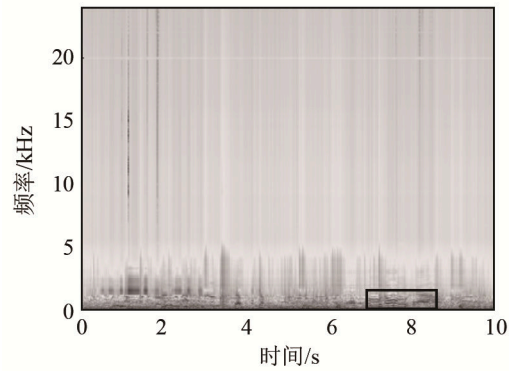


图 6 系数矩阵 $\boldsymbol{H}$

Fig.6 Coefficient matrix $\boldsymbol{H}$



(a) 原始声谱图



(b) NMF 重构声谱图

图 7 原始声谱图和 NMF 重构声谱图

Fig.7 Original spectrogram and NMF reconstructed spectrogram

4 实验结果与分析

采用对比实验的方法，分别验证所提出的 LME+NMF 融合特征和 MCNN 模型对 ASC 任务的有效性，并对实验结果进行分析比较。

4.1 实验数据

表 2 给出了 DCASE2017 和 DCASE2018 数据集包含的声学场景类别及场景数。

在 DCASE2017 的 ASC 任务中，开发集提供了包括 15 个真实声学场景的 4 680 个音频样本。每个场景的样本都记录于不同的地点，每个地点记录 18～25 个样本，所有样本都是采样率为 44.1 kHz、持续时间为 10 s 的双声道音频[15]，分配在 4 个折叠当中。折叠为样本的构成方式，用来做交叉验证。以 4 折交叉验证为例，每次取 1/4 作测试，剩余 3/4 做训练。用 4 次结果的均值来衡量模型的性能好坏。

在 DCASE2018 的 ASC 任务中，开发集提供了包括 10 个真实声学场景的 8 640 个音频样本。这些样本记录于 6 个欧洲城市的不同地点，每个地点记录 30～36 个样本，所有样本都是采样率为 48 kHz、持续时间为 10 s 的双声道音频[4]，全部分配在一个折叠当中。

表 2 实验使用的声学场景数据集

Table 2 Acoustic Scene Datasets in the Experiments

数据集	声学场景	场景数
DCASE 2017	沙滩、公共汽车、咖啡馆/饭店、汽车、市中心、林荫道、便利店、家、图书馆、地铁站、办公室、公园、住宅区、火车、电车	15
DCASE 2018	机场、公共汽车、地铁、地铁站、公园、广场、商场、步行街、公路、电车	10

4.2 基线系统

DCASE2017 的基线系统是一个多层感知机模型，包含两层隐藏单元为 50 的全连接层，隐层和输出层激活函数分别为 Relu 和 Softmax，损失函数为交叉熵，学习率为 0.001，优化器为 Adam[16]。输入样本为单声道 40 维、500 帧(10 s)的 LME 特征图，每帧长 40 ms 且有 50%重叠。

DCASE2018 的基线系统是一个包含两层卷积和两层全连接的 CNN 模型，卷积核大小为 7×7，步长为 1，padding 方式为 same，池化方式为最大池化，同时采用批量标准化和 dropout，其余参数设置和 DCASE2017 基线系统相同。

4.3 结果分析

4.3.1 MCNN 模型效果

为验证 MCNN 模型的效果，用 MCNN 模型分别与 2017 年和 2018 年的基线系统进行对比。实验中除模型外其余设置全部和基线系统相同，结果如图 8 和图 9 所示。

从图 8 和图 9 可以看出，MCNN 模型在 DCASE2017 和 DCASE2018 数据集上都有较好的表现，在大多数声学场景类别上分类精度都有明显提高，在少数几个声学场景类别上和基线系统表现类似，平均分类精度都得到了明显提升，分别提升了 5.9%和 8.6%。

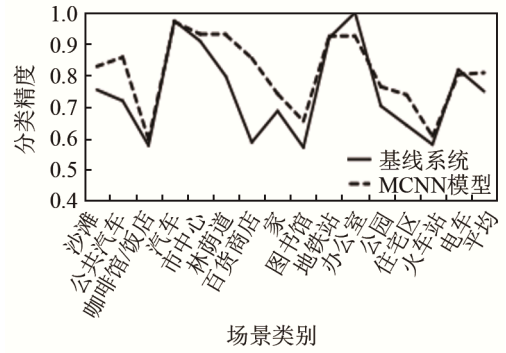


图 8　MCNN 模型与基线系统在 DCASE2017 数据集上的分类精度对比

Fig.8　Comparison of classification accuracy between MCNN model and baseline system on DCASE2017 dataset

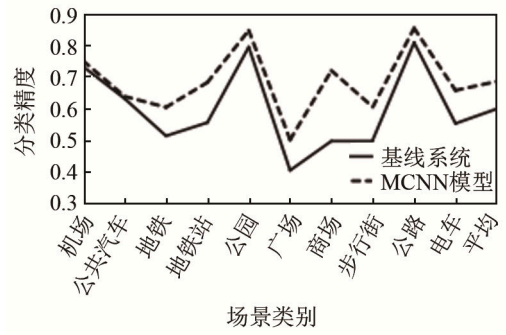


图 9　MCNN 模型与基线系统在 DCASE2018 数据集上的分类精度对比

Fig.9　Comparison of classification accuracy between MCNN model and baseline system on DCASE2018 dataset

图 10 为基线系统和 MCNN 模型在 DCASE2018 数据集上的损失(Loss)曲线，可以看到，MCNN 模型的损失曲线整体都位于基线系统损失曲线下方，说明模型在收敛效果方面要优于基线系统。

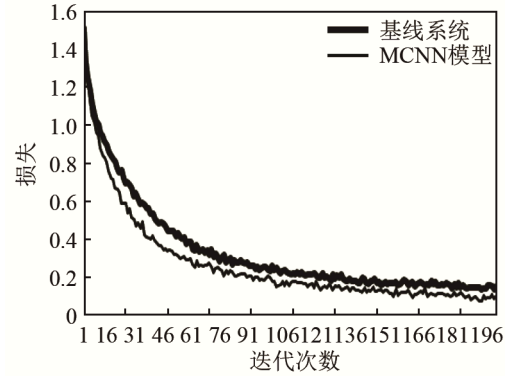


图 10　MCNN 模型与基线系统在 DCASE2018 数据集的损失曲线

Fig.10　Loss curves of MCNN model and baseline system on DCASE2018 dataset

4.3.2　融合特征效果

为验证融合 NMF 特征和 LME 特征对分类效果的影响，在 MCNN 模型下分别使用 40×500 的 LME 特征和 40×500×2 的融合特征进行对比实验，同样采用 DCASE2017 和 DCASE2018 数据集。

对比实验结果如表 3 和表 4 所示。可以看出，在 DCASE2017 数据集上，融合特征在 15 个类别中的 8 个类别上相对 LME 特征精度有所提高；在 DCASE2018 数据集上，融合特征在 10 个类别中的 7 个类别上相对 LME 特征精度有所提高，且平均分类精度分别提高了 1.6%和 1.7%。相比于基线系统，本文提出方案的平均分类精度分别提高了 7.5%和 10.3%，用 DCASE2017 测试集数据进行测试，基线系统精度为 61%，本文提出方案的精度为 68.1%，提高了 7.1%。

表 3　DCASE2017 数据集上的分类结果

Table 3　Classification results on DCASE2017 dataset

场景	分类精度/%	
	LME	LME+NMF
沙滩	82.70	84.60
公共汽车	85.60	96.80
咖啡馆/饭店	59.60	61.20
汽车	96.80	95.80
市中心	92.90	94.60
林荫道	92.60	89.70
百货商店	85.30	83.70
家	74.00	67.90
图书馆	65.40	73.40
地铁站	92.30	90.10
办公室	92.30	98.10
公园	76.30	80.10
住宅区	73.70	68.30
火车	60.60	72.40
电车	80.10	78.20
平均	80.70	82.30

表 4　DCASE2018 数据集上的分类结果

Table 4　Classification results on DCASE2018 dataset

场景	分类精度/%	
	LME	LME+NMF
机场	74.70	75.80
公共汽车	63.60	62.40
地铁	60.20	65.10
地铁站	68	71.80
公园	84.30	86.80
广场	50.00	48.60
商场	71.70	74.90
步行街	60.30	55.90
公路	85.00	87.40
电车	65.50	70.90
平均	68.30	70.00

5 结 论

本文针对声学场景分类任务，提出一种低维度的特征融合机制，融合了低层次的 LME 特征和高层次的 NMF 特征，得到一种既包含整体包络信息又包括细节谐波结构信息的特征；提出适应于声音时频结构的多分辨率特征融合的 MCNN 模型，使得模型可以从特征图中学习到更全面的信息。在 DCASE2017 和 DCASE2018 数据集上的实验结果表明，分类精度得到了显著提升，验证了本文所提方法的有效性。

参 考 文 献

[1] RABINER L R, JUANG B H. Fundamentals of speech recognition[M]. Englewood Cliffs: PTR Prentice Hall, 1993.

[2] DEUTSCH D. Music recognition[J]. Psychological Review, 1969, **76**(3): 300-307.

[3] STOWELL D, GIANNOULIS D, BENETOS E, et al. Detection and classification of acoustic scenes and events[J]. IEEE Transactions on Multimedia, 2015, **17**(10): 1733-1746.

[4] MESAROS A, HEITTOLA T, VIRTANEN T. A multi-device dataset for urban acoustic scene classification[J]. Submitted to DCASE 2018 Workshop, 2018, **2**(1): 1-5.

[5] MESAROS A, HEITTOLA T, DIMENT A, et al. DCASE 2017 challenge setup: Tasks, datasets and baseline system[C]//DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events, 2017.

[6] MESAROS A, HEITTOLA T, VIRTANEN T. TUT database for acoustic scene classification and sound event detection[C]//Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016: 1128-1132.

[7] AUCOUTURIER J J, DEFREVILLE B, PACHET F. The bag-of-frames approach to audio pattern recognition: a sufficient model for urban soundscapes but not for polyphonic music[J]. J. Acoust. Soc. Am., 2007, **122**(2): 881-891.

[8] CAUCHI B. Non-negative matrix factorisation applied to auditory scenes classification[D]. Paris: Université Pierre et Marie Curie, 2011.

[9] ERONEN A J, PELTONEN V T, TUOMI J T, et al. Audio-based context recognition[J]. IEEE Transactions on Audio Speech & Language Processing, 2006, **14**(1): 321-329.

[10] RAKOTOMAMONJY A, GASSO G. Histogram of gradients of time-frequency representations for audio scene classification[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 2015, **23**(1):142-153.

[11] EGHBAL-ZADEH H, LEHNER B, DORFER M, et al. CP-JKU submissions for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks[C]//IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

[12] ZHENG W P, YI J T, XING X T, et al. Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion[C]//Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017). 2017.

[13] HAN Y, PARK J, LEE K. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification[C]//The Detection and Classification of Acoustic Scenes and Events (DCASE), 2017: 1-5.

[14] SAKASHITA Y, AONO M. Acoustic scenes classification by ensemble of spectrograms based on adaptive temporal divisions[C]//Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), 2018.

[15] MESAROS A, HEITTOLA T, VIRTANEN T. TUT database for acoustic scene classification and sound event detection[C]//Signal Processing Conference (EUSIPCO), 2016 24th European. IEEE, 2016: 1128-1132.

[16] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. Arxiv Preprint ArXiv: 1412.6980, 2014.