Computer Science and Application 计算机科学与应用, 2025, 15(3), 180-188
Published Online March 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.153070

Hans汉斯

在多特征下基于卷积神经网络与注意力机制的环境声分类研究

赵乾曜[1]，田益民[2*]，李纪元[1]，孙兆永[2]

[1]北京印刷学院信息工程学院，北京
[2]北京印刷学院基础部，北京

收稿日期：2025年2月23日；录用日期：2025年3月20日；发布日期：2025年3月26日

摘　要

为解决传统城市噪音分类中数据过少而导致模型泛化效果不好，鲁棒性过高，同时传统的噪音特征不能解决关键数据丢失问题导致模型准确率下降。本文提出了一种基于MFCC + GFCC混合特征和噪音语谱图特征的双路卷积模型。该模型首先对噪音数据进行MFCC，GFCC和语谱图变化，提取特征数据，将MFCC和GFCC数据分别进行卷积压缩处理，并在混合后进行分类。对于噪音的语谱图特征进行卷积后，使用注意力机制模块对其各个通道进行权重标记后进行分类，将两路的分类结果进行贝叶斯数值融合，从而实现对城市噪音的正确分类。实验结果表明，识别的准确率比传统模型网络在大数据样本的数据集下有了8%左右以上的提升。

关键词

噪音分类，混合特征，卷积网络，注意力机制

Research on Environmental Sound Classification Based on Convolutional Neural Network and Attention Mechanism under Multiple Features

Qianyao Zhao[1], Yiming Tian[2*], Jiyuan Li[1], Zhaoyong Sun[2]

[1]Faculty of Information Engineering, Beijing Institute of Graphic Communication, Beijing
[2]Foundation Department, Beijing Institute of Graphic Communication, Beijing

Received: Feb. 23rd, 2025; accepted: Mar. 20th, 2025; published: Mar. 26th, 2025

*通讯作者。

文章引用: 赵乾曜, 田益民, 李纪元, 孙兆永. 在多特征下基于卷积神经网络与注意力机制的环境声分类研究[J]. 计算机科学与应用, 2025, 15(3): 180-188. DOI: 10.12677/csa.2025.153070

Abstract

In order to solve the problem of poor generalization effect and high robustness of the model due to too little data in the traditional urban noise classification, the accuracy of the model decreases due to the fact that the traditional noise features cannot solve the problem of key data loss. In this paper, a two-way convolution model based on MFCC + GFCC hybrid features and noise spectral features is proposed. Firstly, the noise data is changed by MFCC, GFCC and spectrogram, the feature data is extracted, the MFCC and GFCC data are convoluted and compressed respectively, and the classification is carried out after mixing. After convoluting the spectral features of noise, the attention mechanism module is used to classify each channel by weighting labeling, and the classification results of the two channels are fused with Bayesian numerical values, so as to achieve the correct classification of urban noise. Experimental results show that the accuracy of recognition is improved by more than 8% compared with the traditional model network under the dataset of big data samples.

Keywords

Noise Classification, Hybrid Features, Convolutional Networks, Attention Mechanisms

1. 引言

近些年来城市现代化发展越来越迅速，对于城市声音识别(Environment Sound Recognition, ESR)领域不断突破，声音识别已经在多个领域，如工业，医疗，民生，语音助手等相关子领域已经研究得相当成熟，而在城市噪音分类的城市管理领域内的研究则相对较少[1]。对于如何利用人工智能来实现城市的智能化管理，已经有很多城市开始进行逐步地研究与探索。城市中的不同来源的噪音使用神经网络来进行智能识别与分类可以帮助更好地管理与控制噪声产生，同时智能城市中各种传感器技术的成熟使实时动态的传播大量的噪音数据，使实时分析成为可能[2]。

对于复杂多变的城市噪音环境，环境声音特征的提取往往直接决定了一个分类模型的好坏，对于声音特征提取，往往分为去除噪音，平滑化，标准化以确保分析的稳定性和准确性，声音被处理成短时窗口，通常使用 STFT 或 MFCC 方法将长时间的声音数据分解为短时频谱片段，对于特征数据处理阶段，有很多人提出了关于特征提取的各种方法。LUZ 等人使用频域特征和时域特征组合的聚合特征，使用时域频域混合特征能更加全面地表达声音在数据上的表示，蔡等[3]提出了一种子带能量规整感知线性预测系数特征，通过减去子带能量偏差来规整时频能量，能有效补偿语音特征信息，噪音环境下模型的平均识别精度提升 68%。除了特征提取之外，一个优秀的模型结构也是对于分类模型结果好坏的关键，卷积神经网络(CNN)模型对于处理分类问题一直都有良好的效果文献[4]中研究了使用 CNN 网络来进行城市噪音分类，文献[5]中还对长短时间记忆网络(LSTM)和门控循环单元(GRU)这两种循环神经网络(RNN)的改进的网络与噪音分类问题做了探讨与研究，除此之外还有一些传统的机器学习算法如感知器随机森林算法也有不错的效果，如 Pilos A 等人[6]提出了一个基于 MFCC 单一特征和多感知器进行分类的声音识别模型，环境声音的不同图像表示(频谱图、MFCC 和 CRP)的分类准确性[7]。同时支持向量机(SVM) [8] [9]、K-最近邻(K-NN) [10]、高斯混合模型(GMM) [11] [12]、隐登马尔可夫模型(HMM) [13]和人工神经网

络(ANN)等机器学习分类器被用于 ESC。深度神经网络(DNN)也用于 ESC。使用的各种 DNN 包括卷积神经网络(CNN) [14]、卷积递归神经网络(CRNN) [15]、深度置信神经网络(DBNN) [16]、张量深度堆叠神经网络(TDSNN) [17]和图像识别网络[18]。CNN 在 ESC 中的首次使用是由 Piczak [19]完成的。

上述方法均在不同的数据库中，有着不错的效果，但由于上述方法中有些公共数据库的数据量较少如 ESC-50 数据库，往往不同标签的声音数据只有 50 多中数据，导致数据的泛化能力不理想，在更实际复杂的城市环境中不能完成应有的效果，经研究在自主采集的数据集(该数据集包含五种常见的城市社会噪音，每类噪音 2000 多条数据样本，总共包含 11,322 条数据)，发现使用混合特征和多注意力模糊后分类模型中，实验效果非常卓越。模型贡献主要分为三个方面。① 探索了在噪音处理中特征提取中使用了语谱图，MFCC (梅尔频率倒谱系数)和 GFCC (梅尔频率组合倒谱系数)组合的混合特征应用在声音识别模型中的有效性和可行性。② 引入了注意力机制，探讨了多种注意力机制的组合对模型分类效果的提升和模型可解释性的研究。③ 基于卷积神经网络提出了一种新的网络结构。

2. 方法

提出的一种 ESC 识别模型，该模型由从噪音信号中的多个特征通道和一个新的卷积神经网络组成。特征提取语谱图，梅尔频率倒谱系数(MFCC)和伽马音频频率倒谱系数(GFCC)。对于分类阶段，提出了一种基于卷积神经网络(CNN)的双流结构，它更适用于音频数据的分类，其中一个流来对语谱图进行处理，另一个对 MFCC 和 GFCC 的混合特征进行处理，最后再将处理结果聚合起来进行决策融合。在双流卷积神经网络结构中添加注意力模块，使分类模型更集中特征模型中的关键部位，对于不同通道的空间位置的特征都有不同程度的强调。

2.1. 特征提取

人类能够正确地分辨出说话人声音的不同，是因为人耳的听觉系统具有很高的复杂度。要使机器正确区分说话人，必须对说话人的声纹进行特征提取，使之成为机器可以区分的特征参数[20]。本文采用 MFCC，GFCC 和语谱图组成三通道特征，不仅提高了模型在各种情景下的泛化能力，提取各种特征使模型对不同的频段的声音都有优秀的分类效果。

2.1.1. MFCC 特征

Mel 频率与人耳所听到的声音的频率的关系可以表示为：

$$\mathrm{Mel}(f)=2595\lg\left(1+\frac{f}{700}\right)$$

MFCC 特征提取的第一步是将语谱信号通过一系列 Mel 滤波器进行滤波处理。这些滤波器的频率划分按照 Mel 频率刻度进行，目的是模拟人类听觉系统对声音频率的感知。然后进行离散余弦变换，最后进行倒谱变换通常是取前几个(本文提取前 13 个通道) DCT 系数，这些倒谱系数即为最终的 MFCC 特征参数。MFCC 的提取流程如图 1 所示。

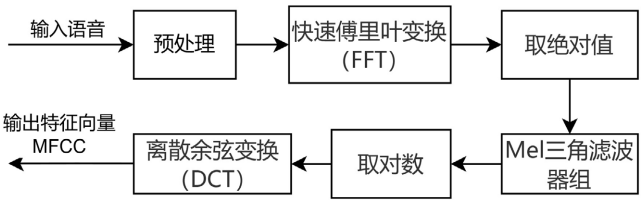

Figure 1. Diagram of the practical teaching system of automation major
图 1. MFCC 提取过程

对于输入的语言信号的预处理主要分为三个部分，预加重：信号通过一个高通滤波器，增强高频部分，减少信号中的噪声和失真。分帧：将预加重后的信号分成短时窗口(通常 20~40 毫秒)，每个窗口称为一帧。加窗：对每一帧应用窗函数(如汉明窗)，以减少帧末端的截断效应。

2.1.2. GFCC 特征

GFCC 特征的提取流程如图 2，输入信号首先通过一组 Gammatone 滤波器。这些滤波器模拟人耳内的基底膜响应，按照非均匀的频率刻度(即伽马音频刻度)对信号进行滤波。每个 Gammatone 滤波器对应于不同的中心频率，能够更准确地反映人类听觉系统对不同频率的敏感性。

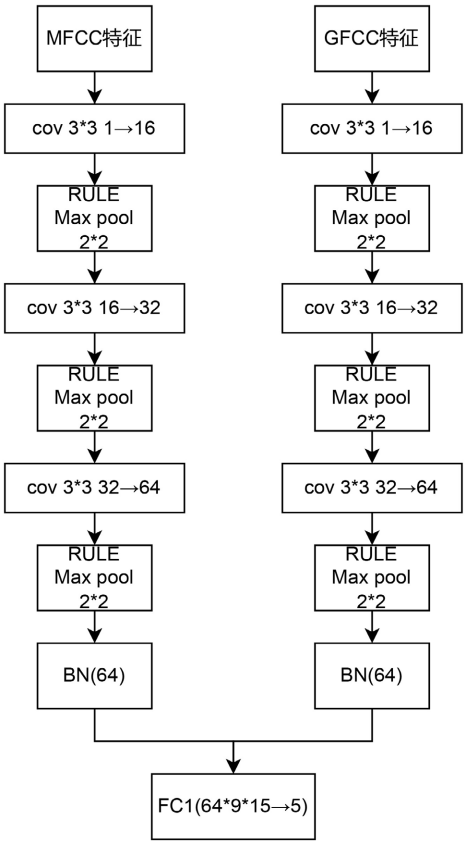

Figure 2. MFCC and GFCC characteristics
图 2. MFCC 和 GFCC 特征

Gammatone 滤波器是一种基于标准耳蜗结构的滤波器，其时域表达式如下：

$$g_i(t) = At^{n-1}e^{-2\pi b_i t}\cos(2\pi f_i + \varnothing_i)U(t),\ t \geqslant 0,\ 1 \leqslant i \leqslant N$$

每个 Gammatone 滤波器的输出是信号在特定频带上的能量或幅度。之后计算每个滤波器输出的功率谱。再对每个滤波器的功率谱取对数，最后对取对数后的功率谱进行离散余弦变换(DCT)，将其转换为倒谱系数。DCT 可以有效地将频谱特征编码为压缩的系数表示。从 DCT 得到的倒谱系数中选取前几个系数，这些系数即为 GFCC 特征。这些特征包含了信号的频谱和时域信息，更接近于人耳听觉系统的响应特性。GFCC 的整个流程强调了对声音频率的更精确模拟，适合于需要捕捉细微听觉特性的任务。

2.1.3. 语谱图

语谱图将信号直接转换成频率和时间的二维图像，本文将信号按 2048 个采样点的窗口进行分割帧移

的长度设置为窗口长度的一半，对每个窗口进行傅里叶变化，取变化后的幅度谱的平方作为该窗口的功率谱，将各个窗口的功率谱合并在一起形成最后的语谱图，语谱图中的每个代表了特定时间的频率范围内信号的能量，其特供了对声音的直观表示，更加全面的提供噪音的全局特征，除此之外 MFCC 和 GFCC 收声音信号质量的影响，例如低信噪比、失真或重叠等问题可能会降低分类的稳健性和准确性，相比 MFCC，语谱图在一定程度上受到信号质量影响的程度可能较小。即使在低信噪比、失真或重叠等情况下，语谱图仍然可以提供较为清晰的频谱信息，有助于从噪音中提取特征。

2.2. 网络结构

本文提出了一种对于噪音分类的新型神经网络结构，该结构为两路卷积神经网络，其中一路网络对 MFCC 特征和 GFCC 特征进行分别处理如图 2 所示，分别包含卷积层，激活层，池化层，归一化层和全连接层。分别将特征数据先后进行三层卷积归一化，之后将两路特征混合之后输入全连接层，另一路卷积网络处理处理语谱图特征如图 3 所示，语谱图特征包含了噪音信号的更多的全局特征，所以这一路神经网络对比另一类网络除了包含卷积层，激活层，池化层，归一化层和全连接层之后，还多了一个注意力模块，使网络可以更从全局的数据中更多的注意到更多的关键数据，语谱图特征先进行一层 5*5 的卷积快，激活层和 5*5 的最大池化层，在进行两层 3*3 的卷积层和 CAM 注意力通道，进行归一化层处理后输入全连接层。

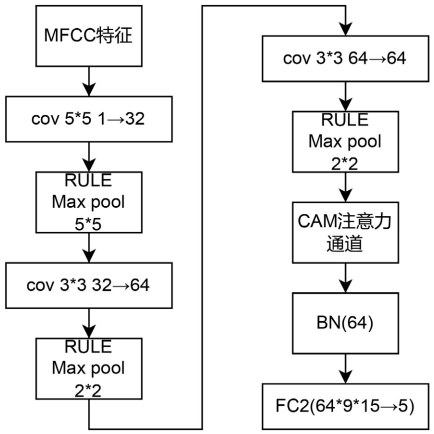

Figure 3. Spectral map features
图 3. 语谱图特征

将两路卷积神经网络处理后的两个全连接层得到的结果，进行平均贝叶斯融合，最后将融合后的结果输出如图 4 所示，最后得到双路卷积神经的分类。

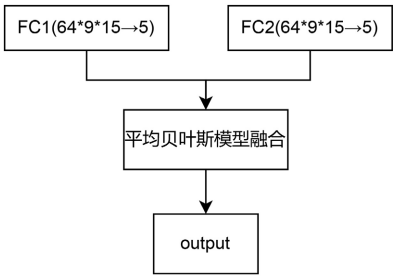

Figure 4. Mean Bayesian fusion
图 4. 平均贝叶斯融合

2.3. 注意力机制模块

注意力机制是一种用于增强特征图中重要通道的表现力的技术，主要用于卷积神经网络(CNN)中。它通过计算每个通道的重要性系数，以自适应地调整特征图中的通道权重，从而提高模型对特征的学习能力[21]。本文中在第二路卷积中使用注意力机制是因为对比 MFCC 和 GFCC 特征，语谱图特征包含了声音信号的全部信息，经过三层卷积之后，语谱图数据为 16*64*9*13 的数据，通过本模型中的注意力模块，模块结构如图 5，该模块使用两个池化操作，分别为平均池化(avg_pool)和最大池化(max_pool)，这两种池化分别从输入特征图中提取通道的全局平均信息和全局最大信息。通过一系列全连接层将池化得到的通道描述向量进行压缩与扩展。经过 ReLU 层激活后数据再恢复到 64 的维度。这一过程能够挖掘出通道之间的关系，并强调重要通道。最终，平均池化和最大池化生成的两个输出向量相加，形成一个通道权重向量。然后，它通过 Sigmoid 函数得到一个范围在[0, 1]的系数，使得每个通道可以根据其重要性进行加权。

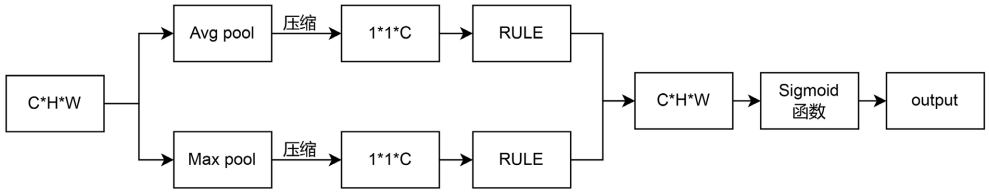

Figure 5. Attention module
图 5. 注意力模块

3. 实验验证

3.1. 数据集

我们所提出的模型在数据集上获得的先进实验结果。数据集中包含 5 个社会声音类别，分别为自然噪音，施工噪音，交通噪音，社会噪音，工业噪音。数据集中包含 11,322 条实验数据，每条实验数据长度为 4 秒钟，总时长为 704 分钟，采样频率为 16 khz。数据集详细信息如表 1 所示，数据集的训练采用五折交叉验证训练和测试模型的性能。

Table 1. Noise data details
表 1. 噪音数据详细信息

数据类别	样本数量	总时长/min	包含声音类别	数据类别
自然噪音	2199	146.6	鸟叫、蝉鸣等动物叫声和雷声雨声等噪音	自然噪音
施工噪音	2097	139.8	挖掘、打洞、搅拌等	施工噪音
交通噪音	2220	148	街道汽车、铁路、城市轨道地铁、机场噪声等	交通噪音
社会噪音	2263	150.8	商场叫卖声，公园广场人流声等噪音	社会噪音
工业噪音	1793	119.3	冲床、打夯、风机等工业机器噪音	工业噪音

3.2. 实验设置

我们的实验基于表 1 数据集。我们采用五折交叉验证训练和测试模型的性能，评估指标包括：

a) 准确率(Accuracy)：模型预测正确的样本比例。

b) 精确率(Precision)：预测为正样本中实际为正样本的比例。

c) 召回率(Recall)：实际为正样本中被预测为正样本的比例。

d) F1-score：精确率与召回率的调和平均数，用于综合评估模型性能。

硬件方面实验在 NVIDIA GTX 4060GPU 上进行，软件方面，所有的实验都在编程环境为 python3.10，模型基于 pytorch 框架进行分类模型的搭建，音频特征提取基于 librosa 和 torchaudio 库。在训练阶段，采用交叉熵函数作为损失函数，使用其衡量模型预测的概率分布与真实标签之间的差异，使用的优化器为 Adam 优化器学习率为 0.001，Adam 优化器结合了动量(Momentum)和自适应学习率(Adaptive Learning Rate)，训练过程中，优化器根据计算出的梯度调整每个参数的值。

3.3. 实验结果分析

在本节中，我们将分析提出的模型在数据集上的性能表现。特别地，我们关注于模型在文本分类任务中的准确率、召回率和 F1-score 等指标。通过与当前主流模型进行比较，我们希望展示我们的方法在处理特定类型数据时的优势。如表 2 所示，本文模型相对比传统的 CNN 模型，RNN 模型，LSTM 模型在本文数据集下各种指标值有明显提升。

Table 2. System resulting data of standard experiment
表 2. 不同模型的性能指标

模型	Accurary	Percision	召回率 Recall	F1-score
CNN	83.42	83.73	83.42	83.48
RNN	81.10	81.68	81.09	81.21
LSTM	83.48	83.73	83.48	83.55
our	92.69	92.71	92.69	92.70

由表 2 可以看出，在语音识别中表现效果很好的循环神经网络(RNN)和长短时间记忆网络(LSTM)对于处理噪音分类问题效果不是很好，处理二维数据更占优势卷积神经网络(CNN)有不错的效果，可以得到对于噪音分类问题卷积神经网络是更好选择，故此本文使用的基本模型框架选用卷积层对噪音特征进行处理，本文使用模型的比其基本的网络模型在本数据集中无论准确率，准确率，召回率 F1-score 都有了大幅度提升大约提升了 10%。同时为了证明混合特征的双路卷积网络是否优于单一特征的通道网络，通过实验对比不同的单路通道网络和本模型对于分类精度的影响。表 3 列出了不同的组合得出对比结果，对于 MFCC，GFCC，MFCC 和 GFCC 混合特征，spectrogram 特征分别都使用本文网络模型中的相对应部分进行训练，来进行对比实验。

Table 3. Model data for different features
表 3. 不同特征的模型数据

特征	模型设置	Accurary	Percision	召回率 Recall
MFCC	AlexNet	83.32	83.54	83.31
GFCC	AlexNet	86.10	86.68	86.10
MFCC + GFCC	AlexNet	90.28	90.30	90.28
spectrogram	SENet	84.79	85.10	84.79
our	SENet	93.21	93.24	93.21
our	our	96.32	96.45	96.32

不同特征的模型数据表示当模型的输入为 MFCC 特征和 GFCC 特征混合时，相比起单一的 MFCC 或者 GFCC 作为模型的输入，模型的准确率有了大约 4%到 6%的提升。由表 3 可以看出，虽然语谱图特征的模型数据不如 MFCC + GFCC 混合特征，但语谱图特征中包含更多的声音信息，故此本文的模型采用的 MFCC + GFCC 混合特征和语谱图特征的双路卷积结合，准确率，准确率，召回率 F1-score 都有了提升，对于声音信息的处理更加全面分类更加精准，语谱图特征的使用能够提供许多对比模拟人耳的 MFCC 和 GFCC 特征中被隐藏的信息。可以从原始的声音数据的层面来使模型的泛化性和鲁棒性提升。

同时因为语谱图数据的数据量偏大，保留了声音频谱图的全部数据，因此使用注意力机制对不同的特征通道分配不同的权重，能够自动突出重要的特征，同时抑制不重要的特征。这种选择性增强有助于模型更好地捕捉关键信息。通过强调重要特征并抑制噪声，通道注意力可能帮助模型在训练过程中降低过拟合的风险，提升其在测试集上的泛化能力。对此通过表 4 可以看出对于语谱图特征，和混合特征，分别进行无注意力模块网络训练，和有注意力机制模块训练。

Table 4. Performance metrics for different models
表 4. 不同模型的性能指标

模型	Accurary	Percision	召回率 Recall	F1-score
CNN	83.42	83.73	83.42	83.48
RNN	81.10	81.68	81.09	81.21
LSTM	83.48	83.73	83.48	83.55
our	92.69	92.71	92.69	92.70

4. 未来展望

城市噪音数据通常是多样且复杂的，包括交通噪音、建筑施工噪音、自然环境声等。获取高质量的噪音数据样本并进行准确标注。以前研究的数据集多为小数据少分类的小型样本，本文采集的数据集包含多种分类，包含一万多条数据，为模型分类研究提供了一个大型样本的研究样本。噪音信号的特征往往比较复杂，模型分类效果往往不理想，本文提出的基于 MFCC 和 GFCC 的混合特征和语谱图组成双路卷积神经网络的输入，同时使用注意力机制模块对语谱图特征全部信息更好的关注。提高了特征表示能力，更好地解决了城市环境噪音中存在大量的干扰声音，这些背景噪音可能会遮蔽目标噪音信号，使得分类任务变得更加困难，同时兼顾人耳对噪音的直观感受和时域频频对噪音的处理。实验结果表明，提出的网络模型对数据集的分类精度达到了 93.69%。有效地提高了环境声音的准确性。

基金项目

本研究得到以下两个项目支持：

Yunfei Du，School of Basic Education，Beijing Institute of Graphic Communication，Beijing 102600，China，项目：the Project of Beijing Municipal Commission of Education (KM 202110015001)；

北京印刷学院重点教学改革项目——工程认证背景下的工科数学教学改革对大学生创新思维与创业能力培养的研究与实践。

参考文献

[1] Muhammad, G., Alotaibi, Y.A., Alsulaiman, M. and Huda, M.N. (2010) Environment Recognition Using Selected MPEG-7 Audio Features and Mel-Frequency Cepstral Coefficients. 2010 5*th International Conference on Digital Telecommunications*, Athens, 13-19 June 2010, 11-16. https://doi.org/10.1109/icdt.2010.10

[2] Luz, J.S., Oliveira, M.C., Araújo, F.H.D. and Magalhães, D.M.V. (2021) Ensemble of Handcrafted and Deep Features for Urban Sound Classification. *Applied Acoustics*, **175**, Article ID: 107819. https://doi.org/10.1016/j.apacoust.2020.107819

[3] 蔡尚, 金鑫, 高盛翔, 等. 用于噪音鲁棒性语音识别的子带能量规整感知线性预测系数[J]. 声学学报, 2012, 37(6): 667-672.

[4] Cao, J., Cao, M., Wang, J., Yin, C., Wang, D. and Vidal, P. (2018) Urban Noise Recognition with Convolutional Neural Network. *Multimedia Tools and Applications*, **78**, 29021-29041. https://doi.org/10.1007/s11042-018-6295-8

[5] 孙陈影, 沈希忠. LSTM 和 GRU 在城市声音分类中的应用[J]. 应用技术学报, 2020, 20(2): 158-164.

[6] Pillos, A., *et al*. (2016) A Real-Time Environmental Sound Recognition System for the Android OS. *Detection and Classification of Acoustic Scenes and Events* 2016, Budapest, 3 September 2016, 1-5.

[7] Boddapati, V., Petef, A., Rasmusson, J. and Lundberg, L. (2017) Classifying Environmental Sounds Using Image Recognition Networks. *Procedia Computer Science*, **112**, 2048-2056. https://doi.org/10.1016/j.procs.2017.08.250

[8] Theodorou, T., Mporas, I. and Fakotakis, N. (2015) Automatic Sound Recognition of Urban Environment Events. 17*th International Conference*, *SPECOM* 2015, Athens, 20-24 September 2015, 129-136. https://doi.org/10.1007/978-3-319-23132-7_16

[9] Zhang, X., Zou, Y. and Shi, W. (2017) Dilated Convolution Neural Network with Leakyrelu for Environmental Sound Classification. 2017 22*nd International Conference on Digital Signal Processing* (*DSP*), London, 23-25 August 2017, 1-5. https://doi.org/10.1109/icdsp.2017.8096153

[10] Chu, S., Narayanan, S., Kuo, C. and Mataric, M. (2006) Where Am I? Scene Recognition for Mobile Robots Using Audio Features. 2006 *IEEE International Conference on Multimedia and Expo*, Toronto, 9-12 July 2006, 885-888. https://doi.org/10.1109/icme.2006.262661

[11] Barchiesi, D., Giannoulis, D., Stowell, D. and Plumbley, M.D. (2015) Acoustic Scene Classification: Classifying Environments from the Sounds They Produce. *IEEE Signal Processing Magazine*, **32**, 16-34. https://doi.org/10.1109/msp.2014.2326181

[12] Muhammad, G., Alotaibi, Y.A., Alsulaiman, M. and Huda, M.N. (2010) Environment Recognition Using Selected MPEG-7 Audio Features and Mel-Frequency Cepstral Coefficients. 2010 5*th International Conference on Digital Telecommunications*, Athens, 13-19 June 2010, 11-16. https://doi.org/10.1109/icdt.2010.10

[13] Bountourakis, V., Vrysis, L. and Papanikolaou, G. (2015) Machine Learning Algorithms for Environmental Sound Recognition: Towards Soundscape Semantics. *Proceedings of the Audio Mostly* 2015 *on Interaction with Sound*, Thessaloniki, 7-9 October 2015, 1-7. https://doi.org/10.1145/2814895.2814905

[14] Mushtaq, Z. and Su, S. (2020) Environmental Sound Classification Using a Regularized Deep Convolutional Neural Network with Data Augmentation. *Applied Acoustics*, **167**, Article ID: 107389. https://doi.org/10.1016/j.apacoust.2020.107389

[15] Sang, J., Park, S. and Lee, J. (2018) Convolutional Recurrent Neural Networks for Urban Sound Classification Using Raw Waveforms. 2018 26*th European Signal Processing Conference* (*EUSIPCO*), Rome, 3-7 September 2018, 2444-2448. https://doi.org/10.23919/eusipco.2018.8553247

[16] Gencoglu, O., Virtanen, T. and Huttunen, H. (2014) Recognition of Acoustic Events Using Deep Neural Networks. 2014 22*nd European Signal Processing Conference* (*EUSIPCO*), Lisbon, 1-5 September 2014, 506-510.

[17] Khamparia, A., Gupta, D., Nguyen, N.G., Khanna, A., Pandey, B. and Tiwari, P. (2019) Sound Classification Using Convolutional Neural Network and Tensor Deep Stacking Network. *IEEE Access*, **7**, 7717-7727. https://doi.org/10.1109/access.2018.2888882

[18] Yao, K., Yang, J., Zhang, X., Zheng, C. and Zeng, X. (2019) Robust Deep Feature Extraction Method for Acoustic Scene Classification. 2019 *IEEE* 19*th International Conference on Communication Technology* (*ICCT*), Xi'an, 16-19 October 2019, 198-202. https://doi.org/10.1109/icct46805.2019.8947252

[19] Piczak, K.J. (2015) Environmental Sound Classification with Convolutional Neural Networks. 2015 *IEEE* 25*th International Workshop on Machine Learning for Signal Processing* (*MLSP*), Boston, 17-20 September 2015, 1-6. https://doi.org/10.1109/mlsp.2015.7324337

[20] 周萍, 沈昊, 郑凯鹏. 基于 MFCC 与 GFCC 混合特征参数的说话人识别[J]. 应用科学学报, 2019, 37(1): 24-32.

[21] Zhang, Z., Xu, S., Zhang, S., Qiao, T. and Cao, S. (2021) Attention Based Convolutional Recurrent Neural Network for Environmental Sound Classification. *Neurocomputing*, **453**, 896-903. https://doi.org/10.1016/j.neucom.2020.08.069