保留立体声相位信息的声音场景分类系统

杨浩聪　史创　李会勇

(电子科技大学信息与通信工程学院，四川成都 611731)

摘　要：针对立体声音频采集设备逐渐普及的趋势，本文提出了一种保留立体声相位信息的声音场景分类算法。在预处理阶段，根据左右通道的相位信息对音频样本进行源环境提取，生成一种全新的四通道特征。在此基础上，集成多个卷积神经网络，搭建一个针对立体声音频录音的声音场景分类系统。区别于现有声音场景分类系统只使用时频谱的幅度信息，本文所提出的方法保留了立体声音频的相位信息。这使得声学特征中所包含的空间方位信息更丰富，立体声音频的优势得到发挥。实验结果证明保留立体声相位信息的声音场景分类系统具有更好的性能，在2019年IEEE音频和声学信号处理技术委员会举办的声音场景分类赛事中相比于基线系统的整体识别准确率提升了18.3%。

关键词：声音场景分类；卷积神经网络；集成学习；源环境提取；双通道相位信息

中图分类号：TN912.16　　**文献标识码：**A　　**DOI:**10.16798/j.issn.1003-0530. ****.**.***

ACOUSTIC SCENE CLASSIFICATION SYSTEM USING BINAURAL PHASE INFORMATION

Yang Haocong　Shi Chuang　Li Huiyong

(School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China)

Abstract: With increasing devices supporting the recording of binaural audios, binaural audio processing methods become a field of possible exploration in acoustic scene classification (ASC). Therefore, we would like to investigate the primary ambient extraction (PAE), a binaural audio processing method which decomposes a binaural audio sample into four channels using the phase information. Features carrying binaural phase information were therefore extracted. An ensemble of convolution neural networks (CNNs) was adopted as the classifier. Compared to existing works, the ASC system proposed in this paper can generate features with additional phase information and make full use of the advantages of binaural audios. The evaluation results validate that the performance of our ASC system can be improved by taking the binaural phase information into account. Our ASC system outperforms the baseline system provide by the 2019 IEEE AASP Challenge Detection and Classification of Acoustic Scenes and Events (DCASE) by 18.3% in terms of the classification accuracy.

Key words: acoustic scene classification, convolutional neural network, ensemble learning, primary ambient extraction, binaural phase information.

1　引言

我们每天会接收到各种声音，并据此来判断我们在哪里或者周边正在发生的事情，这种两种情形分别被称作声音场景和声音事件[1]。就听觉感知而言，声音场景就是从来自不同声源的声音叠加所形成的一个

收稿日期：2020-03-31；修回日期：2020-05-26

基金项目：民航联合研究基金，主动降噪座椅的智能传感理论和控制方法研究，项目编号 U1933127。

复杂声信号中提炼出来的一种高阶知识[2]。声音场景分类会涉及到复数个声源叠加，需要设备听懂自己周围的各种声源并据此给出所处环境的判断。对于已经完成预定义的监督分类，某一特定声音场景类别所包含的声音事件集合也是无界的。同时没有一种分类方式可以覆盖所有的类别，因为理论上声音场景分类的分类数是没有上限的。这些技术难题导致声音场景分类问题成为了机器听觉领域中最困难的任务之一。

自 1997 年 MIT 媒体实验室提交第一份技术报告起[3]，早期的声音场景分类研究就长期受限于计算力水平和模式识别领域发展的瓶颈。当时的研究思路都专注于时间演化特征，因此通常会使用隐马尔可夫模型（Hidden Markov Model, HMM）或者混合高斯模型（Gaussian Mixture Model, GMM）[4][5]，现阶段已经被深度学习方法替代。

2013 年到 2016 年这一期间，随着深度学习的兴起[6]，在计算机视觉和模式识别方法快速演进的促进下[7]，相关研究逐渐出现和深度学习初步结合。彼时声音场景分类领域内的数据集还比较匮乏，因此支持向量机（Support Vector Machine, SVM）仍然是更优先的选择[8]。在这种情况下，来自奥地利约翰开普勒林兹大学计算感知系的 Mesaros 等人开始尝试将深度学习方法加入声音场景分类并初见成效[9]。他们不但尝试使用了卷积神经网络（Convolutional Neural Networks, CNN）作为分类器，同时也将对数梅尔能量谱（Log-mel energies）作为声学特征。由于早期的一些实验心理学研究，梅尔倒谱系数（Mel-scale Frequency Cepstral Coefficients, MFCC）一度被认为是在声音场景分类中最有效的特征。但随后，对数梅尔能量谱逐渐取代了梅尔倒谱系数，卷积神经网络也开始被广泛使用。

2017 年依然还没有具有足够规模的开源数据集供声音场景分类研究使用，只有非公开的 LITIS 数据集初具规模[10]。面对这种困难，Mun 等人使用生成对抗网络（Generative Adversarial Network, GAN）增加训练的数据量，取得了不错的效果[11]。2018 年出现了初具规模的开源数据集[12]，声音场景分类的研究也因此在这一年迎来了爆发。虽然各类研究的效果参差不齐，但是已经普遍倾向于使用卷积神经网络的集成学习这种系统搭建方式。

最近几年声音场景分类取得的研究进展很多，也依然存在着一些问题。虽然研究思路抛弃了时间演化特征，改用时频图与计算机视觉方法相结合。但是目前的时频图特征提取算法无一例外都十分强调幅度信息，而相位信息直接被丢弃。这对于单通道音频而言没有明显的影响，但多通道音频忽略相位信息会造成立体声形成的空间方位信息的丢失，最终使得立体声音频相对于单通道音频不再具备优势。

为了探究相位信息对于声音场景分类性能的影响，我们使用了名为源环境提取（Primary Ambient Extraction, PAE）的非线性立体声音频分离方法[13]，它根据音频左右通道对应的相位信息来分离时频图中的单个时频窗。与之前的方法相比，它生成了具有附加相位信息的特征。在此基础上，我们采用了数据增强、集成学习、4 折交叉验证等常规机器学习方法。本实验的目的一方面是要搭建一个效果优秀的声音场景分类系统，另一方面也欲使用不同特征的性能对比来展现出相位信息的作用。

2 立体声音频的源环境提取算法

源环境提取算法提出的初衷是解决通道格式的音频信号在录制和回放时会发生的设备不匹配问题。在处理立体声录音时，算法会将其分解为四个通道，分别为左侧的源分量和环境分量，以及右侧的源分量和环境分量。分解的结果确保了左右源分量之间的最大相关性、源分量在整个时频域上的最大稀疏性、以及左右环境分量的能量均衡。

假设在立体声信号的每个通道中都有一个源分量和一个环境分量，它们被写成：

$$x_c(t) = p_c(t) + a_c(t) \quad \forall c \in \{0,1\} \tag{1}$$

其中，音频通道索引$c \in \{0,1\}$。源分量p_c被假设是相关的。它们只是振幅不同，通过一个偏移因子可将它们的关系表示为$p_1 = k\,p_0$。环境分量a_c被假设为具有相同的能量但互不相关，而且它们也与源分量不相关。这些都是源环境提取中的关键空间假设。

经过短时傅里叶变换（Short Time Fourier Transform, STFT），（1）式重写为：

$$\boldsymbol{X}_c[m,f]=\boldsymbol{P}_c[m,f]+\boldsymbol{A}_c[m,f]\quad \forall c\in\{0,1\} \tag{2}$$

其中，m是帧的索引，f是频率窗的索引。如图 1 所示，短时傅里叶变换的结果都是复数的，并且可以表示为复数平面中的向量。为了简洁起见，本文后面的部分省略了$[m,f]$。

我们可以将环境分量的短时傅里叶变换时频谱表示为：

$$\boldsymbol{A}_c=|\boldsymbol{A}_c|\odot\boldsymbol{W}_c\quad \forall c\in\{0,1\} \tag{3}$$

在（3）式中，$\boldsymbol{W}_c$指$\boldsymbol{W}_c(m,f)=e^{j\theta_c(m,f)}$，$\theta_c(m,f)$是$\boldsymbol{\theta}_c$位于时频窗$(m,f)$中的元素，而$\boldsymbol{\theta}_c=\angle\boldsymbol{A}_c$是对环境分量相位谱采样所得的向量。

由于$\boldsymbol{P}_1=k\boldsymbol{P}_0$，我们可得：

$$\boldsymbol{X}_1-k\boldsymbol{X}_0=\boldsymbol{A}_1-k\boldsymbol{A}_0 \tag{4}$$

将（4）式代入（3）式：

$$|\boldsymbol{A}|=(\boldsymbol{X}_1-k\boldsymbol{X}_0)/(\boldsymbol{W}_1-k\boldsymbol{W}_0) \tag{5}$$

设向量$(\boldsymbol{X}_1-k\boldsymbol{X}_0)$的相位角为$\boldsymbol{\theta}$，由于$|\boldsymbol{A}|$是非零实数，所以比值关系$\sin\boldsymbol{\theta}/\cos\boldsymbol{\theta}=(\sin\boldsymbol{\theta}_1-\mathrm{k}\sin\boldsymbol{\theta}_0)/(\cos\boldsymbol{\theta}_1-\mathrm{k}\cos\boldsymbol{\theta}_0)$成立，可进一步整理为：

$$\sin(\boldsymbol{\theta}-\boldsymbol{\theta}_0)=k^{-1}\sin(\boldsymbol{\theta}-\boldsymbol{\theta}_1) \tag{6}$$

上式中的$\boldsymbol{\theta}_0$有两种可能的解：

$$\begin{aligned}\boldsymbol{\theta}_0^{(0)}&=\boldsymbol{\theta}-\boldsymbol{\alpha},\\ \boldsymbol{\theta}_0^{(1)}&=\boldsymbol{\theta}+\boldsymbol{\alpha}+\boldsymbol{\pi},\end{aligned} \tag{7}$$

其中，$\boldsymbol{\alpha}=\arcsin[k^{-1}\sin(\boldsymbol{\theta}-\boldsymbol{\theta}_1)]$并且$\boldsymbol{\alpha}\in[-0.5\pi,0.5\pi]$。此外，$(\boldsymbol{W}_1-k\boldsymbol{W}_0)$的虚部与$(\boldsymbol{X}_1-k\boldsymbol{X}_0)$的虚部正负符号相反，导致$(\mathrm{Im}\{\boldsymbol{W}_1-k\boldsymbol{W}_0\}/\mathrm{Im}\{\boldsymbol{X}_1-k\boldsymbol{X}_0\})|_{\boldsymbol{\theta}_0}\geq 0$。由于这个限制条件，最终使得$\boldsymbol{\theta}_0=\boldsymbol{\theta}+\boldsymbol{\alpha}+\boldsymbol{\pi}$是（6）式唯一的解。

将（3）式和（5）式代入（2）式可得：

$$\begin{aligned}&\boldsymbol{A}_c=(\boldsymbol{X}_1-k\boldsymbol{X}_0)/(\boldsymbol{W}_1-k\boldsymbol{W}_0)\odot\boldsymbol{W}_c,\\ &\boldsymbol{P}_c=\boldsymbol{X}_c-(\boldsymbol{X}_1-k\boldsymbol{X}_0)/(\boldsymbol{W}_1-k\boldsymbol{W}_0)\odot\boldsymbol{W}_c,\\ &\forall c\in\{0,1\}\end{aligned} \tag{8}$$

上式中，$\boldsymbol{X}_c$和k可以通过输入信号的相关性计算求得，但是$\boldsymbol{W}_c$仍然未知，进一步讲$\boldsymbol{\theta}_0$和$\boldsymbol{\theta}_1$仍然未知。但由于$\boldsymbol{\theta}_0$和$\boldsymbol{\theta}_1$之间存在相关性，只有一个相位角$\boldsymbol{\theta}_1$需要估计。最终，在稀疏约束下，这个问题可以如下表示：

$$\hat{\boldsymbol{\theta}}_1^{*}=\arg\min_{\hat{\boldsymbol{\theta}}_1}\left\|\hat{\boldsymbol{P}}_1\right\|_1 \tag{9}$$

通过求得源分量的 1 范数最小值，我们可以获取源环境提取的最终结果。

从上述推导过程我们可以看出源环境分离算法以简化的信号模型为基本假设，构建了一个数学上的欠定问题，再使用源分量的稀疏性限定条件确定唯一解。但在这一过程中，无论是信号模型还是限定条件都很难与真实录音场景完全一致，因此源环境分离算法的前期研究工作都通过主观测试来完成有效性验证[14]。

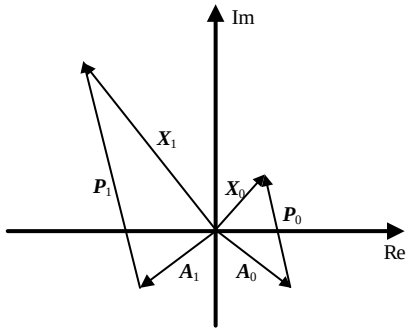

图 1　源环境提取在复数平面上的几何表示

Fig. 1　Geometric representation of PAE in a complex plane

3　声学特征

声音场景分类由于涉及到的场景繁多，导致不同场景的重要特征所处的频带位置差别巨大，也因此使得噪声与特征之间的区别方式非常繁杂微妙。这与频带相对固定却非平稳的语音信号有本质上的区别，这也是为何语音识别领域的研究思路和优秀成果难以快速迁移到声音场景分类领域。如何选取有效的特征至今仍是声音场景分类的一个重要的开放议题。本次实验的声学特征依旧采用音频信号的时频表示，但为了使声学特征更贴近人类的听觉，我们对短时傅里叶变换产生的时频图又进行了一系列尺度转换处理。

3.1　梅尔尺度变换

首先将经过梅尔滤波器组滤波生成的对数梅尔能量谱转化到对数尺度的 dB 域，模拟人耳对响度的辨识度。梅尔尺度是非线性的刻度单位，梅尔（Mel）这个名称源于 melody 一词，表示音阶是基于等距音高比较的。梅尔频率与实际频率 f 的具体关系如下：

$$mel(f)=2595\log_{10}(1+\frac{1}{700}) \tag{10}$$

梅尔频率尺度的对数分布关系更适应于人的感官，在梅尔尺度下我们使用一系列的等距三角形带通滤波器组成梅尔滤波器组，取每个三角形滤波器频率带宽内所有的信号幅度加权和作为该带通滤波器的输出，对时频图进行滤波。滤波器组在普通频域尺度下的分布如图 2 所示。

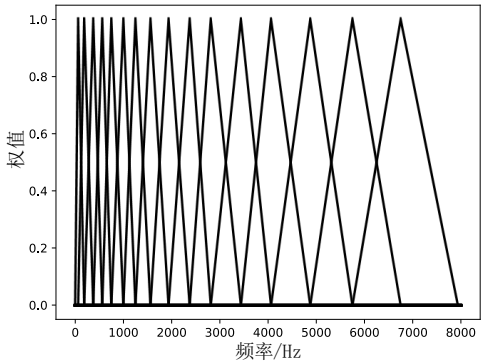

图 2　梅尔滤波器组

Fig. 2　Mel frequency filter bank

3.2　A-weighting 修正

在梅尔滤波的基础上，我们继续进行了 A-weighting 修正，人耳对不同频率音频感知到的相对响度是不同的，频率过高或者过低时感知到的响度都会降低，以 1000 Hz 对应 0 dB 为基准，A-weighting 的修正方式如下：

$$A(f) = 20\log_{10}(R_A(f)) - 20\log_{10}(R_A(1000))$$
$$R_A(f) = \frac{12194^2 f^4}{(f^2+20.6^2)\sqrt{(f^2+107.7^2)(f^2+737.9^2)}(f^2+12194^2)} \tag{11}$$

其中，f 为频率值。通过这些尺度变换和修正，我们让声学特征更符合人耳的响应方式。而变换后的声学特征幅值变化范围也远小于变换前，后续的算法因此可以避免一些数值问题并加快收敛速度。

4 声音场景分类系统搭建

4.1 系统输入

系统输入是长度为 10 s 的双通道立体声录音，对于每个输入的音频样本，我们使用 46 ms 宽的 Hanning 窗和 23 ms 的窗口滑动步进，将其分为 431 帧，之后通过短时傅里叶变换获得对应的时频图。在该时频图的基础上，继续通过窗口数为 128 的梅尔滤波器组，获得对数梅尔能量谱。而后转化到 dB 域进行 A-weighting 修正。最终获得形状为（431,128, n）的声学特征，其中 n 代表通道数。

表 1 针对未裁切特征的卷积神经网络结构

tab. 1 CNN structure for raw features

输入 431×128×1 或 431×128×2 或 431×128×4
7×7 Conv2D (pad=1, stride=1)-32-BN-ReLU 7×7 Conv2D (pad=1, stride=1)-32-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 5×5 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU GlobalAveragePooling2D
Dense (512, activation='relu')
Dense (10, activation='softmax')

通过以上的处理方法生成表示时频信息的特征图后，我们用不同音频通道组合分离分别创建了单通道特征，双通道特征和四通道特征。单通道特征是从单声道音频获得的，它是双声道音频左右通道的几何均值，特征形状为（431,128,1）。虽然单声道音频缺乏空间方位信息，但它的复杂度更低，训练所需的时间开销更低。双通道特征来自于原生录制的双声道音频，左右声道各占用一个通道，特征形状为（431,128,2）。这也是现在多数声音场景分类研究在处理立体声音频时的常用方式，但是在短时傅里叶变换时频图到对数梅尔能量谱转换过程中仅计入了幅度信息，所以造成了左右声道相位信息被忽略。最后，我们使用源环境

提取生成了四通道特征，包括左右通道的源分量和左右通道的环境分量，特征形状为（431,128,4）。我们在短时傅里叶变换后，生成对数梅尔能量谱之前就对时频图使用了源环境提取，相位信息在转化损失前就得到了保留，因为相位信息损失在向对数梅尔能量谱转化的过程中。

为了增强单个模型的泛化性，我们采用了裁切（Cropping）的数据增强方法，裁切通过从现有的特征图中截取子图来生成更多的训练数据[15]。本实验中以 43 帧间隔长度为裁切窗口的滑动步进，从形状为（431,128, n）的对数梅尔能量特征中裁切出 8 个形状为（129,128, *n*）的特征。因此，最终的系统输入种类为 3（通道数）×2（未裁切+裁切）=6 种。

4.2 子分类器网络结构

对于声音场景分类来说，所提取信息更加强调整体的背景声音纹理，而出现在场景中的单个事件片段不太重要。以声音场景中的街道为例，在街道我们会听见车辆行驶的声音，行人的脚步声，嘈杂的谈话声，但是缺失以上一种甚至全部声音事件的街道也是存在的。声音场景中出现的许多事件都不是关键事件甚至是噪声，这也是时频图背景纹理特征取代之前的时间演化特征的原因。而卷积神经网络由于可以识别缩放，移位等空间失真不变性[16]，因此相比于循环神经网络（Recurrent Neural Network, RNN）和长短时记忆网络（Long Short Term Memory Network, LSTM）等序列模型，它在声音场景分类中更具优势。我们使用的卷积神经网络结构如表 1 和表 2 所示。

表 2 针对裁切特征的卷积神经网络结构

tab. 2 CNN structure for cropped features

输入 129×128×1 或 129×128×2 或 129×128×4
3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-32-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-64-BN-ReLU 2×2 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-128-BN-ReLU 5×5 MaxPooling2D
3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU 3×3 Conv2D (pad=1, stride=1)-256-BN-ReLU GlobalAveragePooling2D
Dense (512, activation='relu')
Dense (10, activation='softmax')

这些模型搭建的思路原则遵循于 VGGNet [17]。每个模型包括 1 个输入层，8 个卷积层，1 个全连接层和 1 个输出层。两种模型之间的差异在于前两个卷积层中卷积核的大小，因为裁切前后的特征在尺寸上的差异导致需要两种不同的感受野来保持最终的高维特征有近似的分辨率。所有模型均使用了批归一化（Batch Normalization），它可以通过正则化项来加快学习过程并提高基线水平[18]。

4.3 系统架构

为了进一步提升系统识别准确率的同时又具备更可靠的稳定性，我们对 6 个卷积神经网络子模型进行了集成。如图 3 所示，对数据集施行 4 折交叉验证后，我们训练了多个子分类器并使用它们给出中间预测。中间预测又再次用作随机森林（Random Forest, RF）分类器的输入，该分类器会做出最终决策。此外，我们也对测试样本的中间预测取平均（Averaging），生成另一个决策并进行比较。

5 实验设置

5.1 数据集

本次实验一共涉及到了 3 个数据集，包括 TAU（Tampere University）2019 城市声音场景开发数据集、TAU 2019 城市声音场景排行数据集和 TAU 2019 城市声音场景评估数据集。数据集的采集设备为 Soundman OKM II Klassik/studio A3 和 Zoom F8，在采集过程中麦克风会佩戴在采集者的左右耳上以最大程度还原人类听觉系统的工作方式，这种采集方式也为我们提取相位信息提供了先决条件。其中，TAU 2019 城市声音场景开发数据集是一个高质量的双通道音频数据集，包含了在 10 个欧洲城市中收集的各种声音场景样本。录音的总时长为 40 小时，总共 14400 个 10 s 录音片段，包括机场、巴士、地铁、地铁站、公园、公共广场、购物中心、街道、步行街道、交通电车 10 个分类，每个类别对应 1440 个录音片段。开发数据集包含训练子集和评估子集两部分可以进行初始评估。而 TAU 2019 城市声音场景排行数据集和 TAU 2019 城市声音场景评估数据集均未公开标签信息，仅用于 Kaggle 线上挑战赛在线评估和声音场景分类赛事最终评估。声音场景分类赛事是由 IEEE 音频和声学信号处理技术委员会(AASP)举办的声音场景/事件的检测/分类(Detection and Classification of Acoustic Scenes and Events, DCASE)挑战赛的一个常驻子任务，也是其中历史最悠久，竞争最激烈的子任务。

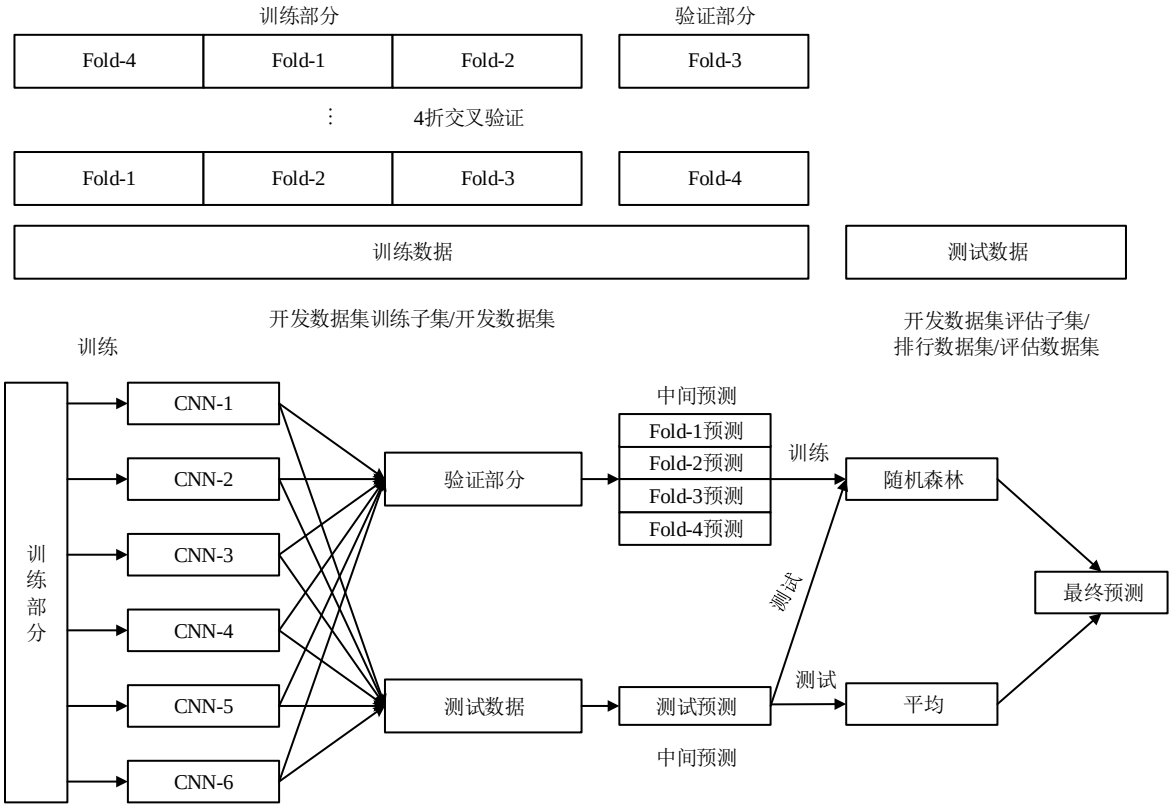

图 3 声音场景分类系统架构

Fig. 3 Architecture of proposed ASC system

5.2 训练配置

首先使用开发数据集中的训练子集对提出的系统进行训练，然后通过评估子集对其进行初步评估。而在参与 Kaggle 线上挑战赛在线评估和 2019 年声音场景分类赛事最终评估时，整个开发数据集都会作为训练数据集训练。在系统输入生成时，我们使用最小最大（Min-max）归一化方法对输入进行修正，以进一步加快收敛速度并避免数值问题。优化器使用随机梯度下降（Stochastic Gradient Descent, SGD）算法，将学习率，衰减和批处理大小分别设置为 0.01、0.0001 和 32。同时，将动量设为 0.9 来加速 SGD 算法。本

实验的硬件为 Intel(R) Core(TM) i5-9600K CPU（3.70GHz）、64 GB 内存、Nvidia GeForce RTX 2070 GPU，软件环境为 Windows10 1809 系统，Python3.6.8、Tensorflow1.8.0、Keras2.2.2。

5.3 损失函数

训练过程中使用多分类交叉熵作为损失函数，表示方式如下：

$$L=\frac{1}{N}\sum_i L_i=\frac{1}{N}\sum_i -\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{12}$$

其中，M 指代分类数量；P_{ic}为观测到的样本i属于类别c的概率；y_{ic}为指示变量，当i=c时其值为 1，否则为 0。

为了使特征超平面中目标优化点附近的采样率更高，使最终优化的落位更接近于最优点，我们使用了混合（Mix-up）方法。它是邻域风险最小化的一种形式[19]，能在数据量不够的情况下进一步填补训练样本之间的空白。它会将两个随机选择的特征以一个权值加权求和生成新的虚拟特征，两者对应的损失函数也进行类似的操作。此过程表示为

$$\begin{aligned}\tilde{x}&=x_i+(1-\lambda)x_j\\ \tilde{L}&=L_i+(1-\lambda)L_j\end{aligned} \tag{13}$$

上式中，x_i和x_j是两个随机选择的特征；L_i和L_j是对应的损失函数。随机变量λ遵循 Beta 分布$Be(\alpha,\alpha)$。当超参数α趋近于零时，模型收敛将趋向回归到经验风险最小化（Empirical Risk Minimization, ERM）[20]。

6 实验结果

6.1 不同特征之间的对比

我们通过相同的方式先后训练了不同的子分类器，它们唯一的差别在输入特征的通道数和感受野的大小，每个子模型通过固定随机种子选取了 4 种不同的训练数据划分方式，最后的平均结果如表 3 所示。这部分结果来自标签完全可知的开发数据集评估子集。

表 3　子分类器识别准确率

tab. 3　The classification accuracies of sub-classifiers

子分类器	全部城市	已知城市	未知城市
CNN-1(单通道)	0.743	0.751	0.663
CNN-2(单通道裁切)	0.771	0.779	0.677
CNN-3(双通道)	0.721	0.731	0.575
CNN-4(双通道裁切)	0.758	0.769	0.630
CNN-5(四通道)	0.734	0.742	0.610
CNN-6(四通道裁切)	0.759	0.771	0.632

以上表中 CNN-2(单通道裁切)为例，它是指将形状为（431,128,1）的单通道特征裁切为（129,128,1）后作为输入训练的卷积神经网络。在开发数据集评估子集中，由于标签完全已知并且每个样本还附带录制地点标签，因此我们对评估子集做了再切分，将其中一个城市录制的样本完全隔离开。这使得训练网络时，管道中未曾出现过来自该城市的录音。该城市对分类器而言是一个未知城市，从而进一步测试网络结构的泛化能力。通过这种分离方式，总共会产生全部城市、已知城市、未知城市三个指标反映模型间的差距。2018 年声音场景分类赛事的组委会明确指出，声音场景分类系统向不同城市迁移也会造成性能下降的情况[12]，以特定场景下识别准确率为优化指标的分类系统不具有可迁移性，因此赛事组委会仅采用整体分类准

确率作为评估指标。为保证与后续实验对比指标的一致性，针对开发数据集评估子集的本地实验也使仅用整体分类准确率来评估。

从表 3 中我们可以看出，无论是裁切前还是裁切后，单个子模型中单通道模型的识别准确率都是最高的，因为单通道输入会减小模型的复杂度并加速收敛，在当前的数据量不足的情况下有天然的优势。所以我们讨论的重点在多通道特征输入模型上。为了遵循单一变量原则，这个实验中有两个关键点我们进行了妥协。首先是我们将每个模型训练迭代次数设置为完全一样，但是复杂度最高的四通道特征在相同迭代次数下收敛到同样的水平是最困难的。此外，源分量和环境分量两者间的特征尺度存在差别。源分量由于更侧重声源，因此对应的关键特征都是时频图上幅值较大并持续一段时间的区域；环境分量更侧重于背景音，因此对应的关键特征大多是时频图上的背景纹理。两者之间的特征尺度差距明显，理论上要用不同大小的卷积窗或者修改网络结构以分别适应两者的感受野，但本次实验已经有特征通道数作为唯一变量，因此未做更多的改变。虽然以上两点的妥协削弱了四通道特征的优势，但最终四通道模型的识别准确率无论是裁切前还是裁切后均超越了双通道模型。这从一定程度上印证了四通道特征比双通道特征更好地保留了立体声录音的相位信息。

6.2 系统性能对比

为了让系统性能的稳定性更具参考意义，我们将各个集成后的系统在三个不同的评估数据集上进行评估。由于 2019 年声音场景分类赛事最终评估的提交限制，我们只有三条验证结果可以与其他参赛系统作比较。具体性能评估见表 4。

表 4　声音场景分类系统识别准确率

tab. 4　The classification accuracies of ASC system

分类系统	开发数据集评估子集			排行数据集(Kaggle)		评估数据集 (2019 年声音场景分类赛事)		
	所有城市	已知城市	未知城市	公开排行榜	个人排行榜	所有城市（95%置信区间）	已知城市	未知城市
基线系统	0.625	0.648	0.438	0.643	0.630	0.633(0.622~0.645)	0.667	0.461
排名中位数系统	/	/	/	0.775	0.765	0.762(0.752~0.772)	0.775	0.696
$RF_{1,2,3,4}$	0.772	0.780	0.682	0.833	0.806	0.812(0.803~0.821)	0.834	0.698
$Averaging_{1,2,3,4,5,6}$	0.778	0.785	**0.693**	0.830	**0.808**	0.799(0.789~0.808)	0.822	0.681
$RF_{1,2,3,4,5,6}$	**0.778**	**0.788**	0.674	**0.840**	0.806	**0.816(0.807~0.825)**	**0.838**	**0.707**

在 Kaggle 线上挑战赛在线评估和声音场景分类赛事最终评估中会使用整个开发数据集来训练，相比开发集训练子集的数据量更大，因此会出现排行数据集和评估数据集的表现优于开发数据集评估子集的情况。从数据量来看，排行数据集的数据量小于开发数据集评估子集小于评估数据集，其中评估数据集（7200 个 10s 录音片段）的数据量更是排行数据集（1200 个 10s 录音片段）的 6 倍，达到完整开发数据集的一半。此外，评估数据集还包含了额外 2 个从未出现在开发数据集中的城市录音，因此无论从数据量还是从样本多样性的角度而言，评估数据集结果的可靠性都要要远高于其余两者，为此我们在评估数据集上额外给出了统计检验结果以进一步反映数据的可靠性。但由于提交次数的限制，我们无法将评估数据集用于前一个对比实验。以表 4 中的 $RF_{1,2,3,4,5,6}$ 为例，它表示用随机森林集成了 6 个子模型，随机森林会把 CNN-1 到 CNN-6 六个子分类器的中间预测作为输入，然后输出最终决策。基线系统由赛事官方提供。排名中位数系统指在 Kaggle 线上挑战赛和声音场景分类最终评估两个排行榜中排名中位数的系统。对于开发数据集我们分别使用了 0，2006，2019 三个随机种子对数据集进行划分并在此基础上进行 4 折交叉验证，表中的结果为各种不同划分下的平均结果。对于评估数据集，声音场景分类赛事的组委会进行了公平统一的第三方评估，最

终结果公正可靠，其中包含了总体识别率以及对应的统计效力。从表中我们可以看出集成了 CNN-5 和 CNN-6 的系统的识别准确率会更高。这也再一次印证了四通道特征比双通道特征更好地保留了立体声录音的相位信息。其中最优秀的提交识别准确率相比基线系统提升了 18.3%，相比中位数排名的系统有 5.4%的提升，而 95%置信区间也小于基线系统和排名中位数系统。

7 结论

本文以搭建立体声声音场景分类系统为出发点，充分考虑立体声录音相比于常规单声道录音的优势来源，对立体声相位信息的作用进行了探究，通过源环境提取方法引入了保留相位信息的四通道特征，获得了一个性能优秀的声音场景分类系统。该系统的分类性能 2019 年 IEEE 声学信号处理技术委员会举办的声音场景分类赛事中排名第四。最终结果一方面获得了声音场景分类赛事组委会的公平统一第三方评估，另一方面也通过一系列本地实验的进行辅助评估。实验对比不同特征输入子模型的性能、不同方式集成学习的性能，印证了四通道特征比双通道特征更好地保留了立体声录音的相位信息，也证明了相位信息的有效性。

参考文献

[1] Mesaros A, Heittola T, Virtanen T. DCASE 2017 challenge setup: tasks, datasets and baseline system[J]. lEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(6): 992-1006.

[2] 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.

[3] Sawhney N, Maes P. Situational awareness from environmental sounds[R]. Boston: Massachussets Institute of Technology, 1997: 1-7.

[4] Eronen A J, Tuomi J T, Klapuri A, et al. Audio-based context awareness-acoustic modeling and perceptual evaluation[C]//Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2003: 529-532.

[5] Eronen A J, Peltonen V T, Tuomi J T, et al. Audio-based context recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(1): 321-329.

[6] Virtanen T, Plumbley M D, Ellis D. Computational Analysis of Sound Scenes and Events[M], Springer, 2018: 3-41.

[7] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[8] Bisot V, Essid S, Richard G. Hog and subband power distribution image features for acoustic scene classification[C]//European Signal Processing Conference, 2015: 724–728.

[9] Mesaros A, Heittola T, Virtanen T. TUT database for acoustic scene classification and sound event detection[C]//European Signal Processing Conference, 2016: 1128-1132.

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

[11] Mun S, Park S, Han D, et al. Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane[R]. Tampere: Tampere University of Technology, 2017: 1-5.

[12] Mesaros A, Heittola T, Virtanen T. A multi-device dataset for urban acoustic scene classification[C]//Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop, 2018: 9-13.

[13] Chen Lu, Shi Chuang, Li Huiyong. Primary ambient extraction for random sign Hilbert filtering decorrelation[C]//International Congress on Acoustics, 2019: 7239-7246.

[14] He Jianjun, Gan W, Tan E. Primary-ambient extraction using ambient phase estimation with a sparsity constraint[J]. IEEE Signal Processing Letters, 2015, 22(8):1127-1131.

[15] Krizhevsky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional neural networks[C]//Neural Information Processing Systems, 2012: 1097-1105.

[16] 胡涛, 张超, 程炳, 等. 卷积神经网络在异常声音识别中的研究[J]. 信号处理, 2018, 34(3), 357-367.
Hu Tao, Zhang Chao, Cheng Bing, et al. Research on abnormal audio event detection based on convolutional neural networks[J]. Journal of Signal Processing, 2018, 34(3), 357-367. (in Chinese)

[17] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition[C]//Computational and Biological Learning Society, 2015: 1–14.

[18] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning, 2015: 448-456.

[19] Zhang Hongyi, Cisse M, Dauphin Y N, et al. mixup: Beyond Empirical Risk Minimization[C]//International Conference on Learning Representations, 2018: 1-13.

[20] Vapnik V, Chervonenkis A Y. On the uniform convergence of relative frequencies of events to their probabilities[J]. Theory of Probability & Its Applications, 1971, 16(2): 264-280.

作者简介

杨浩聪 男，1995 年生，四川绵阳人。电子科技大学信息与通信工程学院硕士研究生，主要研究方向为计算听觉场景分析。
E-mail：yanghaocong@std.uestc.edu.cn

史创 男，1986 年生，辽宁葫芦岛人。电子科技大学信息与通信工程学院副教授，博士学位，主要研究方向为声信号分析与处理，自适应信号处理，线性与非线性声学、噪声及其控制等。
E-mail：shichuang@uestc.edu.cn

李会勇 男，1975 年生，湖北武汉人。电子科技大学信息与通信工程学院教授，博士学位，主要研究方向为阵列信号处理，自适应信号处理。
E-mail：hyli@uestc.edu.cn