引用格式：邓志勇，张万亿，刘爱利．基于二阶差分MFCC深度学习的声景基调声分类方法[J]．中国传媒大学学报（自然科学版），2023，30（05）：26-35+54.
文章编号：1673-4793（2023）05-0026-11

基于二阶差分MFCC深度学习的声景基调声分类方法

邓志勇[1]，张万亿[2]，刘爱利[3*]
（1.首都师范大学音乐学院，北京100048；2.中央音乐学院音乐人工智能与音乐信息科技系，北京100031；3.首都师范大学资源环境与旅游学院，北京100048）

摘要：本文提出了一种可用于卷积神经网络分类技术的二阶差分MFCC特征，尝试解决声景学中基调声与非基调声二分类这一具有“人文色彩”的主观分类任务。以老北京中轴线的声景样本数据集为例，根据本文设计的网络模型结构，使用该二阶差分MFCC特征训练的二分类器对于声景基调声的识别准确率达到80.23%，远优于单独使用RMS和Mel频谱特征，以及联合使用RMS与二阶差分MFCC特征的准确率。
关键词：声景；基调声；卷积神经网络；二阶差分MFCC
中图分类号：O422　**文献标识码**：A

A soundscape keynote classification based on the second order difference MFCC in depth learning

DENG Zhiyong[1], ZHANG Wanyi[2], LIU Aili[3*]
(1. Music College of Capital Normal University, Beijing 100048, China; 2. Department of Music AI and Information Technology, Central Conservatory of Music, Beijing 100031, China; 3. College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China)

Abstract: In order to solve the subjective classification task of soundscape keynote classification with “humanistic color” in depth learning, a feature of the second order difference MFCC used in the classification technology of convolution neural network was put forward in this paper. Taking the soundscape data set in the axis of the Old Beijing for example, the accuracy of the keynote recognition by means of the second order difference MFCC in the designed CNN framework is 80.23%, which is higher than those of RMS, Mel spectrogram, and integration features of RMS and the second order difference MFCC.
Keywords: soundscape; keynote; convolution neural network; second order difference MFCC

1　引言

“声音景观（Soundscape）”，即声景，较早由芬兰社会学与地理学家格兰诺（Johannes Gabriel Granö）在1929年提出，其中心思想是“研究以听者为中心的声音环境”[1]。之后在20世纪六七十代，由加拿大作曲家与生态学家谢弗（R. Murray Schafer）与托阿克斯（Barry Truax）等系统地构建了声景学的理论与方法

基金项目：北京社科基金重点项目（22GLA014）；国家自然科学基金面上项目（41871130）
作者简介（*为通讯作者）：邓志勇（1978-），男，博士，副教授，主要从事声景学与音乐声学研究。Email：dzy@cnu.edu.cn；张万亿（1998-），男，硕士研究生，主要从事音乐科技研究。Email：22tz148@mail.ccom.edu.cn；刘爱利（1981-），男，博士，副教授，主要从事文化地理学与声景学研究。Email：beyondtour@163.com

论框架，认为声景是一种“强调个体或社会感知和理解方式的声音生态”[2]。声景思想提出的初衷是培养人们聆听与改善声音环境的能力，“人-声-境（Human-Sound-Context）”是声景的三个基本层次[3]。2014年，国际标准化组织（ISO）将声景定义为“在某种场境下，由个人或群体感知、经历和（或）理解的声学环境”[4]。

越来越多的音频分类算法结合了不同的特征，例如隐马尔可夫模型、支持向量机、高斯混合模型[5]和本文将使用的卷积神经网络[6]。然而，与语音等音频信号相比，声景样本分类任务具有其特殊性，因为它们具有更广泛的能量频率分布，并且包含更多丰富的人与环境信息，使得其比语音信号更为复杂[7]。基于“人-声-境”的相互关系，谢弗认为声景可分为基调声（Keynote）、信号声（Signal）和坐标声（Soundmark）三种基本类型，指出人们“通常不会有意识的感知基调声”，信号声“与基调声形成对比”，而坐标声是“引起特别考虑与注意的声音”[8]。基于上述观点，本文对声景样本进行分类时，在总体上将声景进一步划分为“基调声（Keynote）”与“非基调声（Non-Keynote）”两个相互正交的类型，如表1所示。

在单个既定的声景样本中，相对于非基调声，基调声往往作为一种声压级动态较小、频率分布较广、时域变化较小的稳态背景声出现，主观听感也相对稳定[9]。例如，图1和图2分别为样本No.71中的长时稳态基调声（于北京景山录制，包含了自然声和游客的喧闹声）和样本No.115中的短时瞬态非基调声（于北京鼓楼录制的击鼓表演时的击鼓声，具体样本编号所对应的区域如表2所示）的光谱图。两者对比可知，图1直观地显示出了典型基调声的稳态特性。

表1 基于谢弗声景类型的基调声与非基调声二分类

基调声		“指某一特定社会中不断或频繁地被听到并足以形成其他声音感知的背景声……通常不会有意识的感知基调声……”
非基调声	信号声	“直接引起特别注意的任何声音……与基调声形成对比……”
	坐标声	“指社区中独特的或拥有的特质能够引起社区民众特别考虑与注意的声音。”

因此，本文在与传统音频分类方法中所采用的均方根包络（Root Mean Square Envelope, RMS）、Mel频谱、梅尔倒谱系数（Mel-frequency Cepstral Coefficients, MFCC）进行比较分析的基础上，尝试提出一种改进后的二阶差分39维梅尔倒谱系数（以下简称“二阶差分MFCC”），以卷积神经网络（Convolutional Neural Network，CNN）对声景样本进行“基调声-非基调声”二分类，构建一个较为完整的深度学习网络模型，以适应此类注重主观听感选择并具有“人文色彩（Humanistic Color）”特征[10]的分类任务。

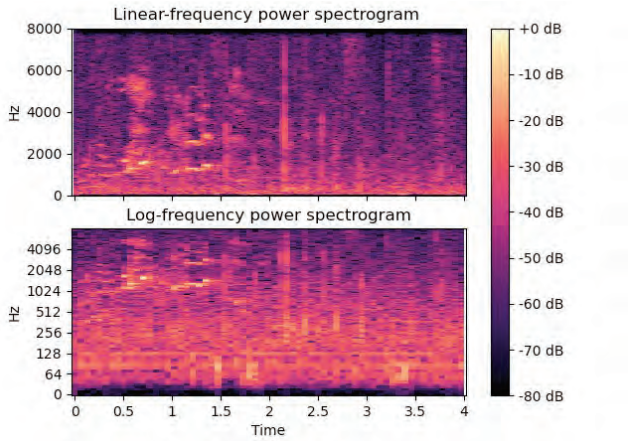


图1 样本No.71中基调声光谱图

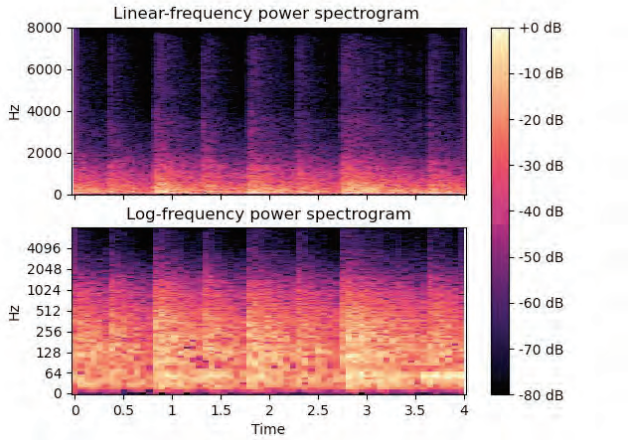


图2 样本No.115中非基调声的光谱图

2 研究样本与方法

2.1 研究样本

本文研究所用的声景样本数据集覆盖了图3红色虚线框所覆盖的老北京中轴线的主要区域，从其南端的永定门至北端的钟鼓楼，总长度约为7.8公里，分为永定门、前门大街、天安门广场、故宫、景山、前海和钟鼓楼及其主要连接道路及周边等七个区域。所有声景样本均为WAV格式，采样频率为48kHz，量化深度为24bit，由专业录音师佩戴Sennheiser Ambeo录音耳机以符合国际标准化组织（ISO）技术标准的双耳全景声制式，以“定点录音（Location Recording）”和“声景漫步（Soundwalk）”两种方式进行录制[11-12]。具体各样

本编号及其所对应的区域,以及各区域中以上两种录制方式的样本编号、数量和录制年份如表2所示。

图3 老北京中轴线声景数据集分布的七个区域

老北京中轴线是一个功能多元、活动繁忙的城市综合区域,交通与人流量大,区域中众多的绿地、公园与知名景点吸引了大量游客与市民。绿地公园的自然声与城市交通噪声及嘈杂的人声不仅是其声景基调声的主要内容之一,同时也提供了大量的非基调声。老北京中轴线上时常会有一些传统音乐演出,如前门大街的叫卖吟唱,景山公园票友们的京戏演唱交流,前海火德真君庙的道教音乐表演,钟鼓楼的鼓乐队演出等,这些音乐声成为了具有"人文色彩"的非基调声。本文所选取的115个样本完整涵盖了图3所标注的老北京中轴线上的七个典型地理区域。这些内容丰富[13-14]的声景样本,在全面体现"基调声-非基调声"二分类的同时,也完整涵盖了声景学研究中常用的包括"自然声(Natural Sounds)"、"人声(Human Sounds)"、"社会声(Sounds and Society)"、"机械声(Mechanical Sounds)"、"安静与沉默(Quiet and Silence)"、"指示声(Sounds as Indicators)"等在内的六个基于内容划分的声音类型[8]。由这些样本所构成的声景数据库,充分满足了深度学习对样本多样性和典型性的要求。

表2 老北京中轴线声景数据集的基本信息

区域名称及其编号	样本数量及其编号	定点样本数量	声景漫步数量	录制年份
1.永定门	13 No.1-13	8	5	2008 2021
2.前门大街	13 No.14-26	8	5	2021
3.天安门广场	8 No.27-34	8	0	2017 2019 2021
4.故宫	35 No.35-69	24	11	2017 2019 2021
5.景山	14 No.70-83	11	3	2009 2011 2021
6.前海	20 No.84-103	6	14	2016 2017 2018 2021
7.钟鼓楼	12 No.104-115	10	2	2021

2.2 数据预处理

根据深度学习模型的样本归一化要求,以表1的描述为标准,表2中的115个样本以"专家评分法(Expert Evaluation)"[15]进行人工分类标注,被分段截取划分为"基调声"与"非基调声"两个正交的层,形成1519个基调声段和1899个非基调声段。然后经过分层随机抽样,形成训练集的样本段总数为2394个,验证集的样本段总数为513个,测试集的样本段总数为511个。即从基调声段和非基调声段中分别随机抽取1453个和941个样本段,形成2394个训练集的样本段;从基调声段和非基调声段中分别随机抽取310个和203个样本段,形成513个验证集的样本段;从基调声段和非基调声段中分别随机抽取306个和205个样本段,形成511个测试集的样本段。最后由于CNN要求可处理的样本数据长度必须保持一致,因此每个样本段最终被划分为等时长的4秒片段,少于4秒的样本段将以零填充尾部至4秒[6]。

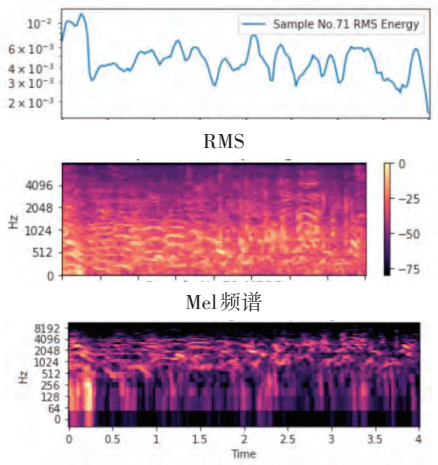
2.3 二阶差分MFCC特征的选取

一维时域的均方根振幅包络(RMS),类似频率非线性特征的Mel频谱和如式(1)的从Mel频谱中获取的带有高冗余信息的梅尔倒谱系数(MFCC)都是音频信号处理中常用的特征[16]。

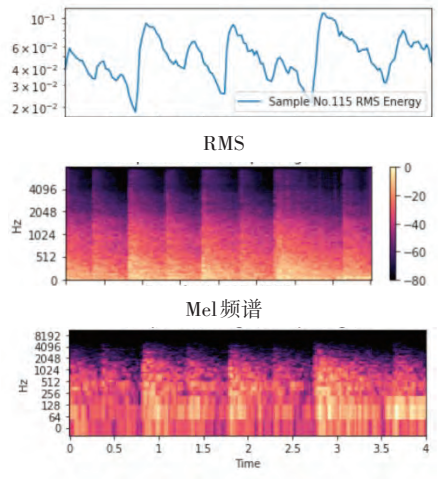
$$\mathrm{MFCC}(i,n)=\sum_{m=1}^{M}\log\left[H(i,m)\right]\cdot\cos\left[\frac{\pi\cdot n\cdot(2m-1)}{2M}\right],n=1,2,\cdots,L \tag{1}$$

其中，i为帧序数；n为列序数；m为Mel频率；$H(i,m)$为能量谱与Mel滤波器转置矩阵的乘积；L为MFCC的维数，本文中L=13；M为Mel滤波器的个数，本文中M=128。

如前所述，由于本文的分类任务具有“人文色彩”的特殊性，因此对式(1)的13维MFCC进行常规二阶差分运算，形成了冗余信息更为丰富的二阶差分39维MFCC(简称为二阶差分MFCC)作为本文的分类特征。图4显示了样本No.71中的基调声和样本No.115中的非基调声的RMS、Mel频谱和二阶差分MFCC的谱图及其直观上的差异。



(a)样本No.71中的基调声



(b)样本No.115中的非基调声

图4 样本No.71中的基调声与样本No.115中的非基调声的RMS、Mel频谱和二阶差分MFCC谱图

2.4 本文CNN结构

根据上节所选取的二阶差分MFCC的分类，并便于与RMS和Mel频谱特征的分类结果进行比较，本文研究的技术路线如图5所示。

在本文设计的分类器中，将基调声和非基调声称为“Class(类)”，每个经人工分类标注划分的样本段称为“Sample(样本)”，对应的类以“Label(标签)”进行标识，数值1表示该样本段为基调声，数值0表示该样本段为非基调声，标签经过“独热编码(one-hot coding)”进行预处理。由于所有深度学习系统中的基本数据结构都为以NumPy数组形式存储的张量(Tenser)，而神经网络的所有输入和目标必须以浮点型张量的形式呈现，且必须对数据进行向量化，因此本文将输入特征数据结构重构为一个包含数值的三维张量，设计的卷积神经网络结构如表3和图6所示。

表3 本文卷积神经网络结构

Layer(type)	Output Shape	Param #
conv2d_1(Conv2D)	(None,126,39,32)	320
conv2d_2(Conv2D)	(None,126,39,32)	9248
conv2d_3(Conv2D)	(None,126,39,32)	9248
conv2d_4(Conv2D)	(None,126,39,64)	18496
max_pooling2d_1(MaxPooling2)	(None,63,19,64)	0
dropout_1(Dropout)	(None,63,19,64)	0
conv2d_5(Conv2D)	(None,63,19,64)	36928
conv2d_6(Conv2D)	(None,63,19,64)	36928
conv2d_7(Conv2D)	(None,63,19,128)	73856
max_pooling2d_2(MaxPooling2)	(None,31,9,128)	0
dropout_2(Dropout)	(None,31,9,128)	0
reshape_1(Reshape)	(None,32,1116)	0
dense_1(Dense)	(None,32,200)	223400
dense_2(Dense)	(None,32,100)	20100
flatten_1(Flatten)	(None,3200)	0
dense_3(Dense)	(None,2)	6402

Total params: 434,926
Trainable params: 434,926
Non-trainable params: 0

在训练过程中，卷积核分别设置为32、64、128个，大小为3*3，将表3中dropout的值设为0.25。不同特征在CNN中具有不同的维度，如表4所示。整个深度学习过程在pytorch框架下实现[6]。

表4 特征阶数

特征	RMS	Mel频谱	二阶差分MFCC
形状	(126,1)	(128,126,1)	(126,39,1)
数据类型	Tenser	Tenser	Tenser
数据维度	1	126	39

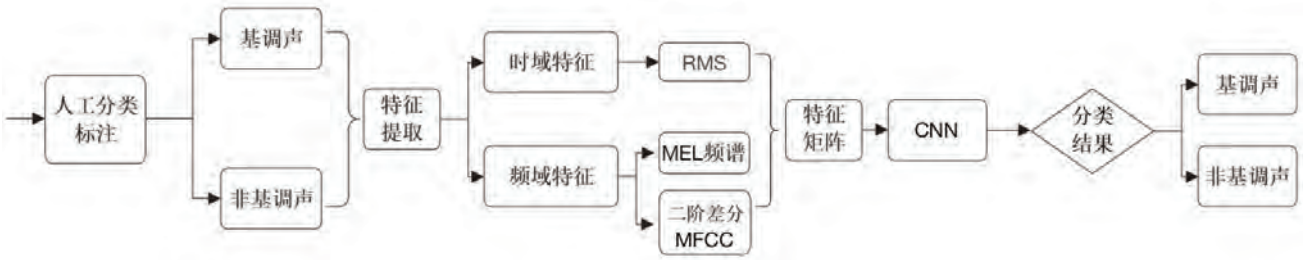


图5 研究路线图

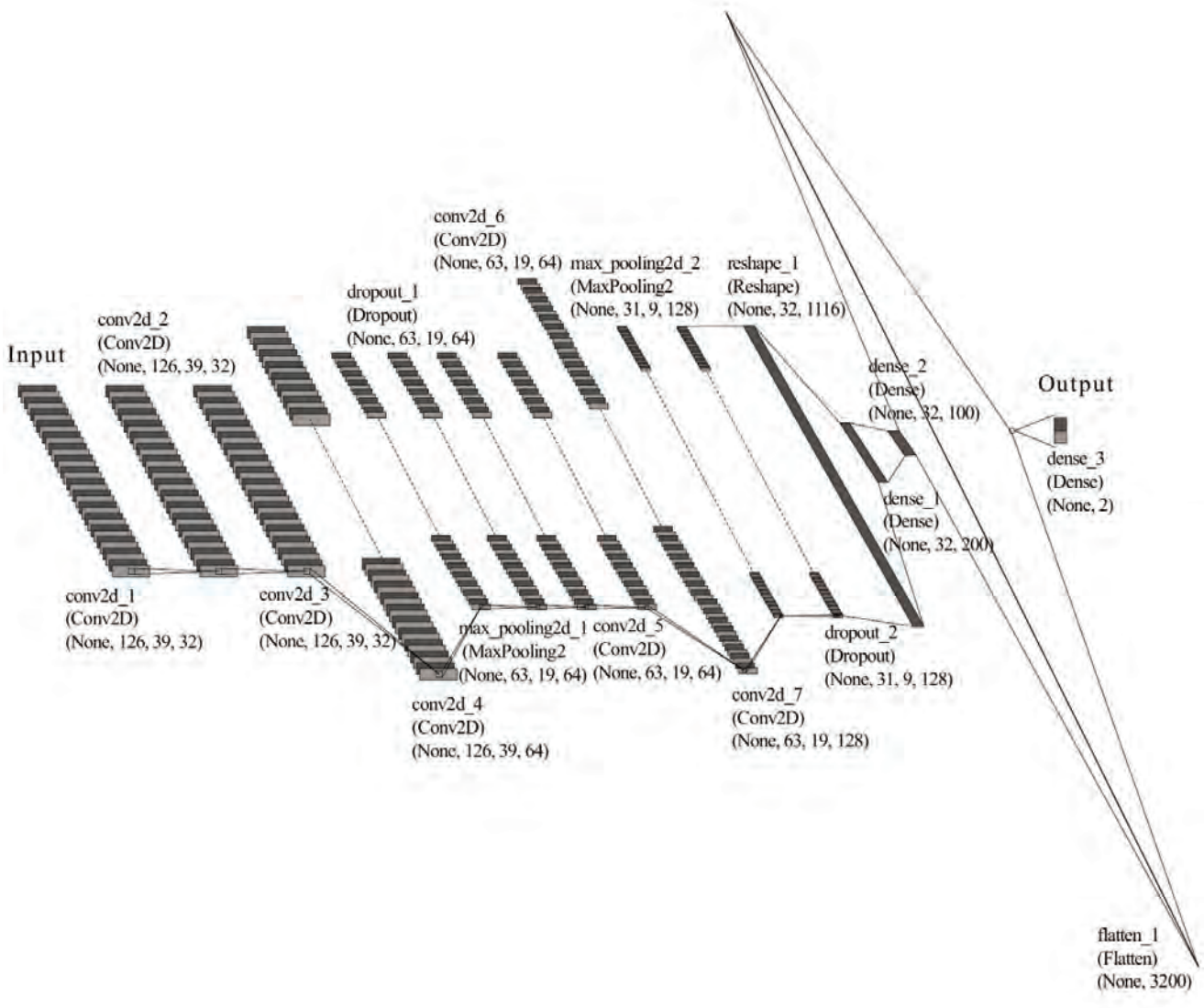


图6 卷积神经网络结构图

2.5 中间层的激活函数与结果验证

由于本文以标签数值标量0和1进行二分类的标识，因此还将选择修正线性单元函数(ReLU函数)作为中间层的激活函数[6]，将上述网络神经元的输入层映射至输出层，以方便进行结果验证。该函数具有将所有负值返回为零的特性，可使网络稀疏，在一定程度上缓解了过拟合，对于二分类结果具有良好的验证性能。ReLU函数的表达式如式(2)，图像如图7所示，经由该函数的神经元输入至输出的加权过程如式(3)所示：

$$y = \mathrm{ReLU}(x) = \begin{cases} x, x > 0 \\ 0, x \leqslant 0 \end{cases} \tag{2}$$

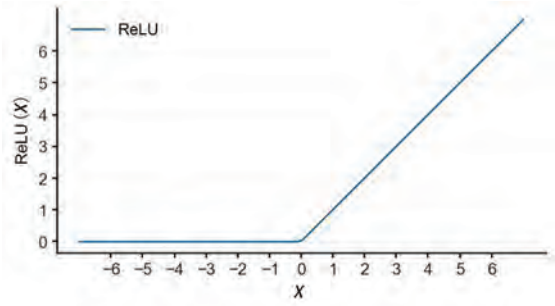


图7 ReLU激活函数的图像

$$y = \mathrm{ReLU}(w_1x_1 + w_2x_2 + w_3x_3 + b) \quad (3)$$

其中，b为偏置量，w_1、w_2、w_3为加权系数。

3 分类结果

3.1 二分类训练的进度控制

表2的数据集经预处理后，使用上述设计的分类器进行训练的部分进度控制如表5所示。

表5 二分类训练的进度控制表

1372/2394[==>	.....]	-ETA:	33s-Loss:	0.6886-acc:	8.5598
1376/2394[==>	.....]	-ETA:	33s-Loss:	0.6886-acc:	8.5596
1380/2394[==>	.....]	-ETA:	33s-Loss:	0.6887-acc:	8.5594
1384/2394[==>	.....]	-ETA:	33s-Loss:	0.6884-acc:	8.5592
1388/2394[==>	.....]	-ETA:	33s-Loss:	0.6886-acc:	8.5584
1392/2394[==>	.....]	-ETA:	33s-1oss:	0.6887-acc:	8.5575
1396/2394[==>	.....]	-ETA:	33s-Loss:	0.6885-acc:	8.5573
1400/2394[==>	.....]	-ETA:	32s-Loss:	0.6886-acc:	8.5571
1404/2394[==>	.....]	-ETA:	32s-Loss:	0.6887-acc:	8.5563
1408/2394[==>	.....]	-ETA:	32s-Loss:	0.6886-acc:	8.5568
1412/2394[==>	.....]	-ETA:	32s-Loss:	0.6886-acc:	8.5574
1416/2394[==>	.....]	-ETA:	32s·Loss:	0.6886-acc:	8.5572
1420/2394[==>	.....]	-ETA:	32s-Loss:	0.6882-acc:	8.5577
1424/2394[==>	.....]	-ETA:	32s·Loss:	0.6883·acc:	8.5569
1428/2394[==>	.....]	-ETA:	32s-Loss:	0.6888-acc:	8.5567
1432/2394[==>	.....]	-ETA:	31s-Loss:	0.6887-acc:	8.5573
1436/2394[==>	.....]	-ETA:	31s-Loss:	0.6888·acc:	8.5564
1448/2394[==>	.....]	-ETA:	31s·Loss:	8.6884·acc:	8.5563
1444/2394[==>	.....]	-ETA:	31s-Loss:	0.6884-acc:	8.5561
1448/2394[==>	.....]	-ETA:	31s-Loss:	0.6884-acc:	8.5559
1452/2394[==>	.....]	-ETA:	31s-Loss:	0.6884-acc:	8.5558
1456/2394[==>	.....]	-ETA:	31s-1oss:	0.6883-acc:	8.5563
1460/2394[==>	.....]	-ETA:	31s-Loss:	0.6881-acc:	8.5568
1464/2394[==>	.....]	-ETA:	30s-Loss:	0.6881-acc:	8.5574
1468/2394[==>	.....]	-ETA:	30s-Loss:	0.6881-acc:	8.5572

该表用于监控深度学习模型训练进度，其中“1372/2394”表示当前训练的“批次数/总批次数”，如在该例中，表示已经完成了1372个批次的训练，总共需要训练2394个批次。“ETA:33s”表示预计剩余完成训练所需要的时间，“Loss:0.6886”表示训练过程中的损失值，“acc:8.5598”则是以百分比表示的训练过程中的准确率。该表提供了当前深度学习模型训练的进度信息，包括已完成的批次数、训练进度的可视化、预计剩余时间、损失值和准确率等指标。通过观察这些指标，可以了解模型训练的进展情况和性能表现。

3.2 输出层的激活函数、损失函数、优化器与二分类识别准确率

为适应二分类任务的特点，本文接下来选择SoftMax函数[6]作为输出层的激活函数，其表达式如式(4)：

$$y = \mathrm{SoftMax}(a_k) = \frac{exp(a_k)}{\sum_{i=1}^{n} exp(a_k)} \quad (4)$$

其中，a_k为向量a的第k个分量。

如果输出神经元的数量为2，则SoftMax函数具有一个属性，即输出值的总和为1，即满足约束条件：$\boldsymbol{P}(\boldsymbol{A}|\boldsymbol{x})+\boldsymbol{P}(\boldsymbol{B}|\boldsymbol{x})=1$。因此，可将SoftMax函数作为一种基于概率统计的方法对目标进行分类。在本文中，输出越接近1，则识别结果与训练集中的基调声相似度越高，反之亦然。

此外，损失函数(Loss函数)[6]是一种二元交叉熵函数，适用指示二分类问题中的分类误差，其表达式如式(5)：

$$\mathrm{Loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1-y_i)\cdot\log(1-p(y_i)) \quad (5)$$

其中，y_i为分类标签：0为非基调声，1为基调声；$p(y)$为某一分类的概率值。

RMSprop函数[6]则是本文选择的优化器，其表达式如式(6)：

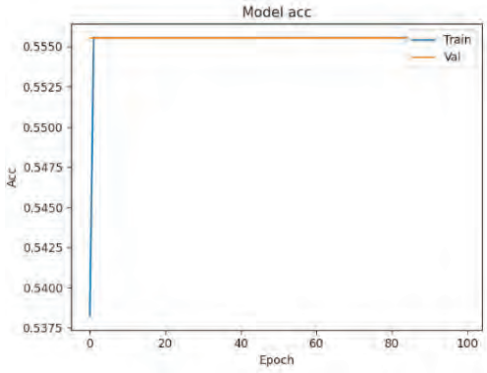
$$\begin{cases} E\left[g^2\right]_t = \alpha E\left[g^2\right]_{t-1} + (1-\alpha)g_t^2 \\ W_{t+1} = W_t - \dfrac{\eta_0}{\sqrt{E\left[g^2\right]_t + \epsilon}} \odot g_t \end{cases} \quad (6)$$

其中，g_t为时间步t的梯度；$E[g^2]_t$为时间步t的梯度平方的移动平均值；α是忘记因子，取值区间为[0,1)，常见值为0.9或0.99；η_0是全局学习率；ε为极小常数，以防止分母为零；W_t是时间步t的参数值。

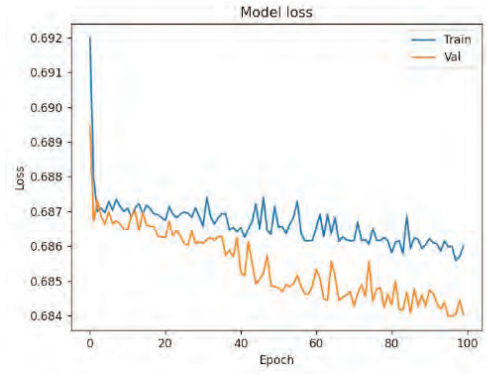
最终，基于2.4节的CNN模型，分别使用RMS、Mel频谱和二阶差分MFCC三种特征进行训练的分类器，在测试集中获得的基调声分类识别准确率如表6所示，三种特征的识别准确率曲线与损失率曲线分别如图8、图9与图10所示。

表6 基调声识别准确率

特征	准确率
RMS	55.56%
Mel频谱	68.68%
二阶差分MFCC	80.23%

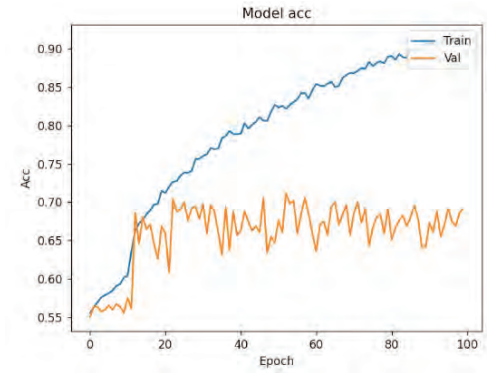


(a) 准确率曲线

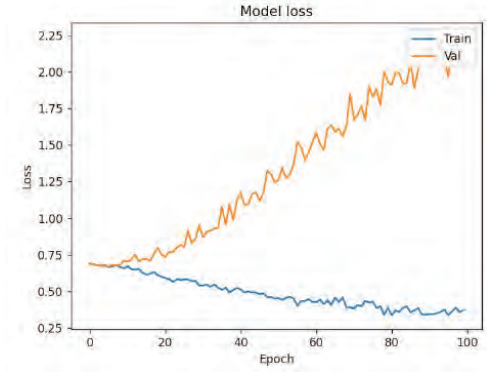


(b) 损失率曲线

图8 使用RMS特征的识别准确率曲与损失率曲线

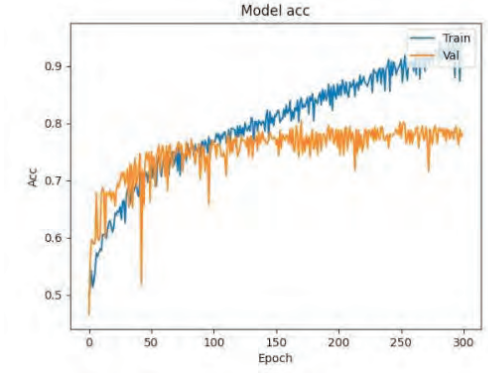


(a) 准确率曲线

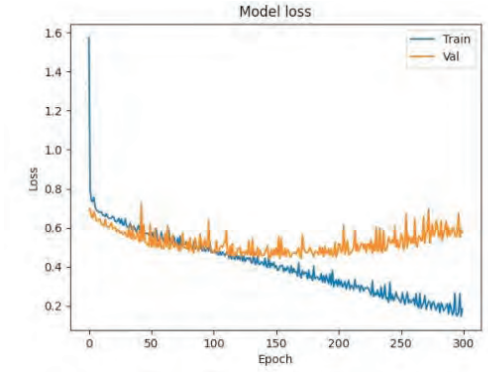


(b) 损失率曲线

图9 使用Mel频谱特征的识别准确率曲与损失率曲线



(a) 准确率曲线



(b) 损失率曲线

图10 使用二阶差分MFCC特征的识别准确率曲线与损失率曲线

由以上分析可知,使用RMS特征的识别结果基本上相当于随机猜测,使用本文提出的二阶MFFC特征的识别结果为80.23%,远高于使用传统Mel频谱特征68.68%的识别结果。

4 评价

4.1 性能评价

本文采用混淆矩阵(Confusion Matrix)[6]对上述分类结果进行性能评价。混淆矩阵为一个两行两列矩阵,每一列的数值表示所识别类别中样本段的数量,每列总和为所识别类别中的样本段总数,每一行则表示样本段的真实属性类别,每行总和则为该类别的样本段总数。因此,混淆矩阵由以下四个指标组成:

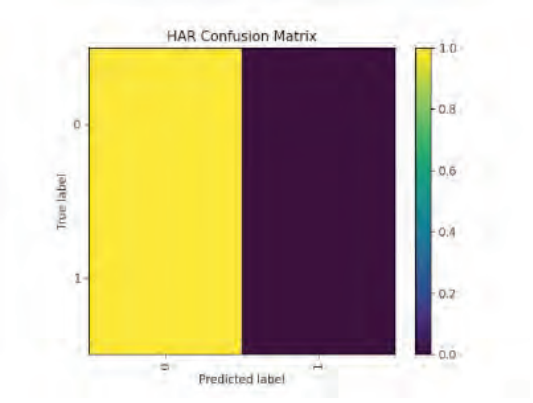
·TP(True Positive),真正例:模型将样本识别为基调声的真实基调声类别样本段数量。

·FN(False Negative),误负例:模型将真实基调声类别样本识别为非基调声的样本段数量。

·FP(False Positive),误正例:模型将真实非基调声类别样本识别为基调声的样本段数量。

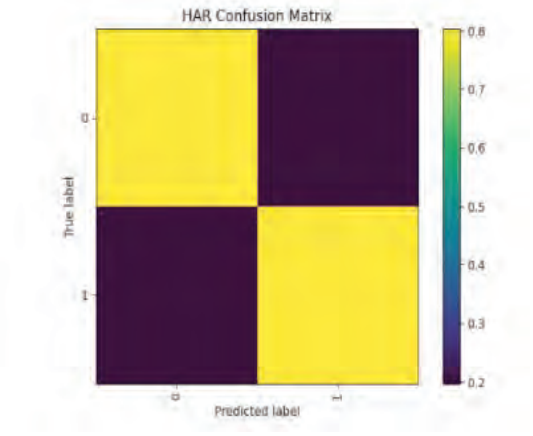
·TN(True Negative),真负例:模型将样本识别为非基调声的真实非基调声类别样本段数量。

分别使用RMS、Mel频谱和二阶差分MFCC三种特征进行训练的分类器混淆矩阵分别如图11、图12和图13所示。



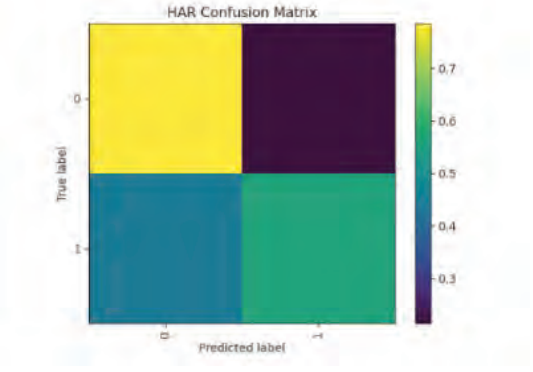
TP = 284，FP = 0，FN = 227，TN = 0

图11 使用RMS特征的混淆矩阵



TP = 223，FP = 61，FN = 99，TN = 128

图12 使用Mel频谱的混淆矩阵



TP = 228，FP = 56，FN = 45，TN = 182

图13 使用二阶差分MFCC的混淆矩阵

可由TP、TN、FP和FN计算出其他四个更为直观的指标来评价二分类模型的性能，如表7所示。四个指标的定义如下：

准确率：Accuracy=(TP+TN)/(TP+TN+FP+FN)，即模型正确识别的样本数量占总样本数量的比例。

精确率：Precision=TP/(TP+FP)，即模型被正确识别为正样本的数量与实际正样本数量的比例。

召回率：Recall=TP/(TP+FN)，即正确识别的样本数量在实际样本数量中所占的比例。

调和度：F1=2*Precision*Recall/(Precision+Recall)，为Precision和Recall的调和平均数。

表7 三种特征进行训练的分类器性能

特征/指标	准确率	精确率	召回率	调和度
RMS	0.5556	1.0000	0.5558	0.7145
Mel频谱	0.6868	0.7852	0.6925	0.7359
二阶差分MFCC	0.8023	0.8082	0.8352	0.8215

4.2 识别能力评价

为了评价分类器识别能力的强弱，需要根据混淆矩阵计算另一个指标ROC曲线[6]。这一曲线的横轴为误正率(False Positive Rate, FPR)，即4.1节中误正例FP的占比，是在二元分类中所有实际负例中被错判为正例的比值；纵轴为真正率(True Positive Rate, TPR)，即表6中的召回率。ROC曲线远离对角线，越趋近于坐标(0，1)时，则分类器模型的整体识别能力越强；ROC曲线越接近于对角线时，识别能力越弱；当ROC曲线为对角线时，则该二分类器模型为无效的随机猜测。参数AUC则表示曲线下的面积：当AUC=1时，是理想的二分类器模型；当AUC=0.5时，则该二分类器模型为无效的随机猜测。AUC值越接近1，则模型的整体识别能力越强。分别使用RMS、Mel频谱和二阶差分MFCC三种特征进行训练的分类器的ROC曲线分别如图14、图15和图16所示。

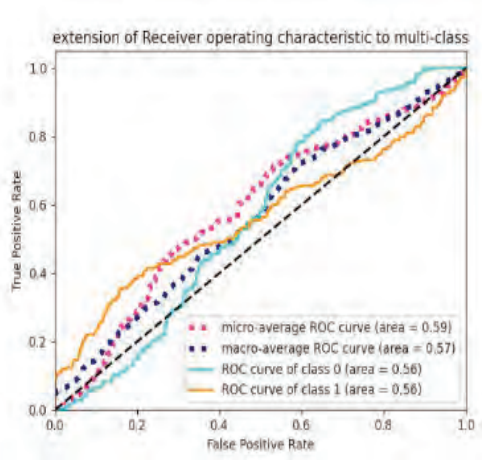


图14 使用RMS特征的ROC曲线

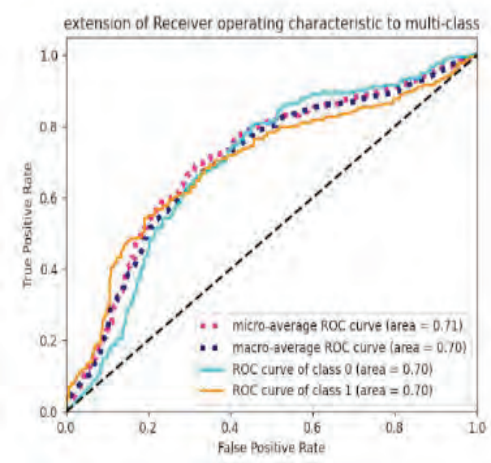


图15 使用Mel频谱的ROC曲线

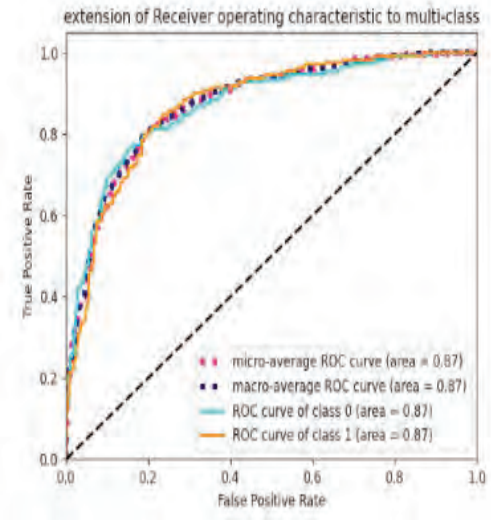


图16 使用二阶差分MFCC的ROC曲线

由ROC曲线可知,RMS的识别能力最弱,接近随机猜测,而本文提出的二阶差分MFCC的识别能力最强。这可能再次表明,像RMS这样的一维时域特征不适用于CNN网络的深度学习模型。不过ROC曲线通常对表示正负样本比例的大幅变化不敏感,因此在本文中,还采用测试数据的得分作为ROC的阈值,以表6中的召回率(Recall)作为横坐标,精确率(Precision)作为纵坐标,生成PR曲线来评价模型的大幅变化特征,分别如图17、图18和图19所示。

PR曲线绘制了在不同概率阈值下,模型的精确率和召回率之间的变化。曲线越接近坐标(1,1),意味着模型在保持高精确率的同时,能够具有高召回率。也就是说,模型能够正确地识别出正例,并且较少将负例误分类为正例。

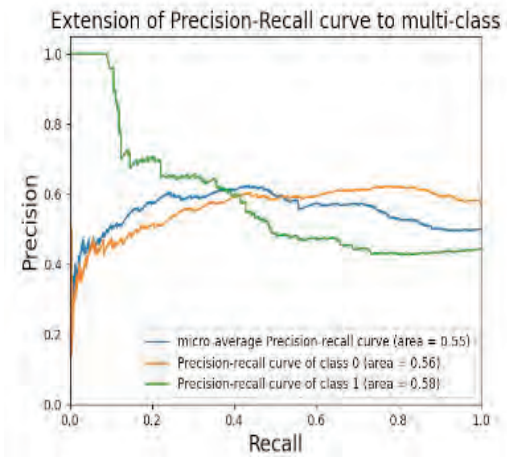


图17 使用RMS特征的PR曲线

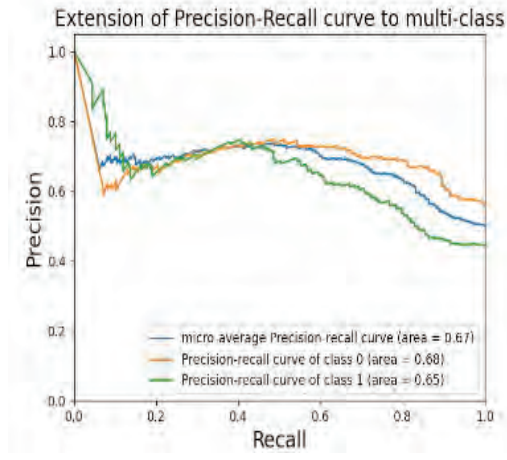


图18 使用Mel频谱的PR曲线

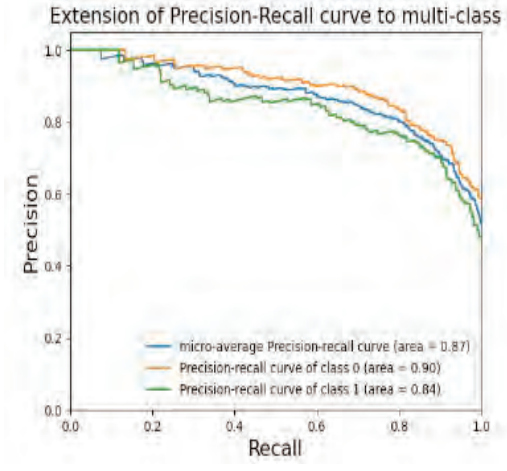
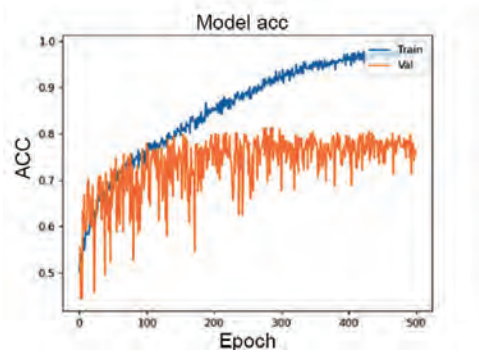


图19 使用二阶差分MFCC的PR曲线

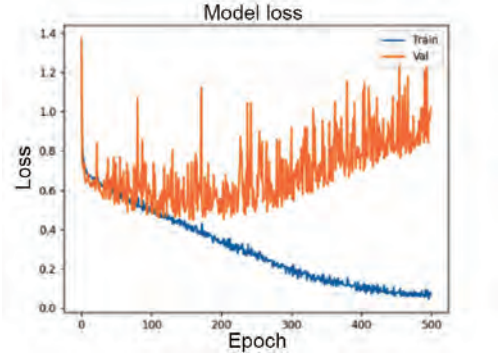
4.3 联合特征的性能评价

如上所述,在CNN网络中,冗余量更为丰富的二

阶差分MFCC比RMS和Mel频谱特征更适用于基调声分类这种具有“人文色彩”的主观分类任务。同时，由于联合特征主要以不同维度的特征信息，通过增加通道来实现更好的性能。因此接下来本文尝试使用RMS和二阶差分MFCC的联合特征，即总共40维信息来对基调声二分类器进行训练，以考察这一联合特征是否能够提高分类器的识别能力。图20、图21、图22和图23分别为使用该联合特征得到的准确率曲线及损失率曲线、混淆矩阵、ROC曲线和PR曲线。



(a) 准确率曲线



(b) 损失率曲线

图20　使用RMS与二阶差分MFCC联合特征的识别准确率曲线与损失率曲线

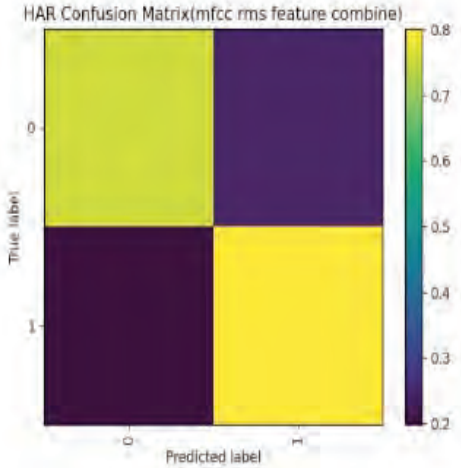


图21　使用RMS与二阶差分MFCC联合特征的混淆矩阵

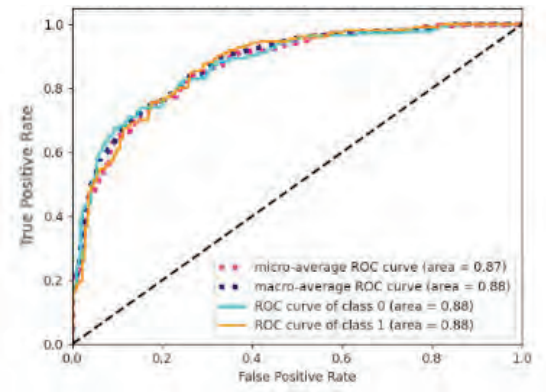


图22　使用RMS与二阶差分MFCC联合特征的ROC曲线

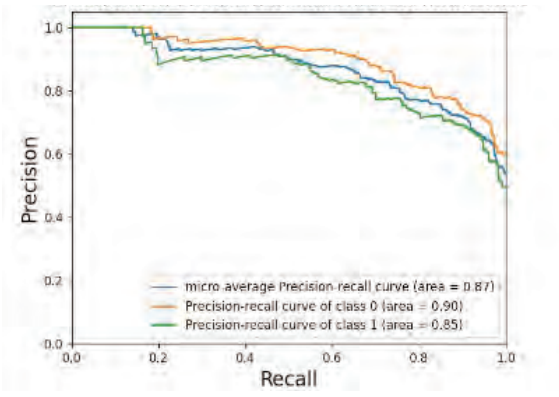


图23　使用RMS与二阶差分MFCC联合特征的PR曲线

由混淆矩阵和上述曲线可知，二阶差分MFCC特征联合RMS并没有提高模型识别的准确率，甚至导致准确率下降了0.78%。

5　结论

综上所述，基于本文提出的二阶差分MFCC特征及网络模型结构进行声景基调声识别的准确率为80.23%，其表现优于单独使用RMS和Mel频谱特征的结果，也优于RMS和二阶差分MFCC特征的联合使用的结果。作为语音识别中最常用的特征之一，结合本文的数据分析结果来看，MFCC及其改进的高维特征仍然适用于声景样本的分类；而基于单一能量特征的一维时域RMS特征可能不适用于声景这类复杂声音的分类任务。

本文仅局限于以老北京中轴线的声景样本数据集为例，针对声景基调声与非基调声的二分类任务，基于CNN网络深度学习的要求，提出了一种在该分类任务中表现良好的高维二阶差分MFCC特征，并与单独使用RMS和Mel频谱特征，以及使用RMS和二阶差分MFCC联合特征，进行了性能上的初步比较。而对于声景这类具有“人文色彩”的复杂声音的主观

（下转第54页）