DOI: 10.11992/tis.201703008
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20170704.1702.010.html

用 Bark 频谱投影识别低信噪比动物声音

黄鸿锂, 李应
（福州大学 数学与计算机科学学院, 福建 福州 350116）

摘　要: 复杂环境声影响低信噪比动物声音的自动识别。为解决这一问题, 本文提出一种不同声场景下低信噪比动物声音识别的方法。该方法把声音信号进行 Bark 尺度的小波包分解, 再使用分解系数生成重构信号的频谱, 并对频谱进行投影生成 Bark 频谱投影特征, 通过随机森林分类器实现低信噪比动物声音的识别。该文分别在流水声环境、公路环境、风声环境和嘈杂说话声环境下, 以不同的信噪比, 对 40 种动物声音进行识别实验。结果表明, 结合短时谱估计法、Bark 频谱投影特征和随机森林的方法对不同信噪比的各种环境声音中动物声音的平均识别率可以达到 80.5%, 且在–10 dB 的情况下依然保持平均 60% 以上的识别率。

关键词: 声音信号; 自动识别; 小波包变换; 随机森林; 环境声音
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)04−0610−09

中文引用格式: 黄鸿锂, 李应. 用 Bark 频谱投影识别低信噪比动物声音[J]. 智能系统学报, 2018, 13(4): 610–618.
英文引用格式: HUANG Hongkeng, LI Ying. Identifying low-SNR animal sounds based on Bark spectral projection[J]. CAAI transactions on intelligent systems, 2018, 13(4): 610–618.

Identifying low-SNR animal sounds based on Bark spectral projection

HUANG Hongkeng, LI Ying
(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)

Abstract: In this paper, we consider the influence of complex background environments on the automatic recognition of animal sounds with low signal-to-noise ratios (SNRs). We propose a method for identifying low-SNR animal sounds in various background environments. In this method, the sound signal is decomposed by a Bark scale wavelet packet, and the decomposition coefficient is used to generate a spectrogram of the reconstructed signal, which is projected onto a spectrogram to generate a Bark spectral projection (BSP) feature. Random forests (RF) are then used to identify animal sounds with low SNRs. We classified 40 common animal sounds with different SNRs in noise environments such as flowing water, highway, wind, and loud speech. The experimental results show that by combining the proposed methods of short-time spectrum estimation, BSP, and RF in various background environments with different SNRs, the mean identification rate for animal noises can reach 80.5%. In addition, a recognition rate above 60% can be maintained even at –10 dB.

Keywords: sound signal; automatic recognition; wavelet packet transform; random forests; environment sound

动物声音自动识别, 对于动物物种、种群及数量研究, 生态环境分析具有重要意义。目前, 对动物声音识别方法的研究有基于时间序列特征的动物声音识别[1], 通过各个音节延续的隐马尔可夫模型的鸟类识别[2], 通过声音模式对鸟类分类[3]。此外, 还有借助于经典的基于文本数据库查询方法, 采用基于索引的动物声音检索[4]以及在连续和真实的现场录音中, 识别特定的鸟类声音[5]。我们也在近年的工作[6-7]中, 通过自适应能量检测进行鸟类声音检测; 对声谱图提取灰度共生矩阵特征, 并结合随机森林 (random forests, RF)[8]识别鸟类声音。然而, 对于自然环境下的各种低信噪比动物声音的识别, 还缺乏有效的方法。

收稿日期: 2017−03−08.　**网络出版日期:** 2017−07−04.
基金项目: 国家自然科学基金项目 (61075022); 福建省自然科学基金项目 (2018J01793).
通信作者: 李应. E-mail: fj_liying@fzu.edu.cn.

关于低信噪比声音信号的分析、分类和识别, 近期的研究包括小波包过滤的低信噪比声音事件识别[9]; 利用匹配追踪 (matching pursuit, MP) 算法从 Gabor 字典中选择重要的原子, 用主成分分析和线性判别分析确定声音事件的特征, 最后采用支持向量机 (supported vector machine, SVM) 分类器进行分类识别[10]。以声谱图及其相关的特征为基础, Dennis 等[11]提出基于声谱图进行伪着色并提取相关图像特征的声音事件识别方法。尤其, Dennis 等[12]提出的子带功率分布 (subband power distribution, SPD) 特征, 在谱图中将可靠的声音事件与噪声分开并去除不可靠区域, 最后用最近邻居分类器 (k-nearest neighbor, kNN) 对特征进行识别。这种方法能在信噪比低至 0 时, 可以识别相关的声音事件。然而, 由于环境声的多变性, 对于自然环境中 0 以下各种低信噪比动物声音及声音事件, 目前却没有更为有效的识别方法。

针对低信噪比动物声音及声音事件的识别, 以声谱图投影特征结合 RF 的动物声音识别方法[13]为基础, 该文提出一种 Bark 尺度的小波包分解系数重构的频谱投影 (Bark scale Wavelet packet decomposition coefficient reconstructed spectral projection, BSP) 特征。并通过结合短时谱估计、BSP 特征和 RF 的方法识别各种声音场景下的动物声音。

1　基于 BSP 的动物声音识别方法

1.1　动物声音识别架构

基于 BSP 特征结合随机森林 (RF) 的动物声音识别的整体架构, 如图 1 所示。具体流程包括: 首先, 对动物声音进行声音增强; 然后将增强后声音信号进行 Bark 尺度的小波包分解并重构分解系数, 把这些重构通过短时傅里叶变换生成重构信号频谱; 并对频谱进行主成分分析, 提取投影特征, 即各个 Bark 频率群的 BSP; 最后使用 RF 识别各个 Bark 频率群的 BSP。

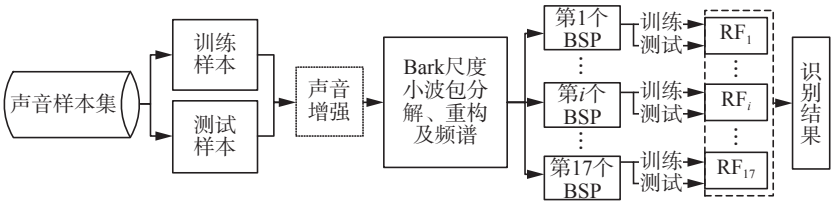


图 1　动物声音识别的流程

Fig. 1　The process of animal sound recognition

1.2　Bark 尺度小波包分解

Bark 是一种模拟人耳听觉感知特性的非线性频率尺度。小波包分析对信号的低频和高频部分同时进行分解, 具有更强的频带划分能力[14]。Bark 尺度小波包分解是基于人耳 Bark 域频率感知特性的小波包分解结构。

人耳的 Bark 域在 20 Hz ~ 16 kHz 的频率范围内分为 24 个 Bark 频率群[15]。Bark 域频率 z 和赫兹 (Herz) 域频率 f 的转换关系为

$$z=\begin{cases}0.01f, & 0<f<500\\ 0.007f+1.5, & 500\leqslant f<1\,220\\ 6\ln f-32.6, & f\geqslant 1\,220\end{cases} \tag{1}$$

式中: Bark 频率群的带宽在 500 Hz 以下时增加速度恒定, 约 100 Hz 增加一个带宽; 在 500 ~ 1 220 Hz 带宽呈线性增加; 1 220 Hz 以上, 带宽呈对数增加。根据小波包分析的特性, 可以用小波包分析来逼近人耳的 Bark 谱。对于 8 kHz 采样、频率在 4 kHz 以下的大部分的动物声音事件, 用常规方法模拟 1 ~ 17 号 Bark, 可以得到图 2, 每个子带的中心频率相差约为 1 Bark 的小波包分解结构。对动物声音识别的第一步, 将按这个分解结构, 对声音信号进行小波包分解。并把这个小波包分解的 17 组系数用于下一步的投影特征提取。

1.3　BSP 特征

声音信号经过小波包分解后, 再对其相应的小波包分解系数重构的频谱进行主成分分析, 得到 BSP 特征。对分解系数重构的频谱投影, 即提取 BSP 特征的过程如下。

1) 计算规范化的频谱矩阵 $\boldsymbol{X}$。对小波包分解系数进行重构, 并把重构的信号进行短时傅里叶变换, 得到重构信号的频谱 $S(t,f)$。其中, t 代表帧索引, $t=0,1,\cdots,M-1$, f 代表频率索引, $f=0,1,\cdots,N-1$。将 S 第 t 帧 $\bar{\boldsymbol{S}}_t=[S(t,0)\ S(t,1)\ \cdots\ S(t,N-1)]^{\mathrm{T}}$, 转化为规范化的帧:

$$\boldsymbol{S}_t=\frac{\bar{\boldsymbol{S}}_t}{\|\boldsymbol{S}\|} \tag{2}$$

$$\boldsymbol{X}=[\boldsymbol{S}_1\ \cdots\ \boldsymbol{S}_t\ \cdots\ \boldsymbol{S}_M]^{\mathrm{T}},\boldsymbol{X}\in R^{M\times N} \tag{3}$$

2) 对频谱矩阵 $\boldsymbol{X}$ 进行特征值分解。$\boldsymbol{C}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, $\boldsymbol{C}=\boldsymbol{U}\Lambda\boldsymbol{U}^{\mathrm{T}}$, 即

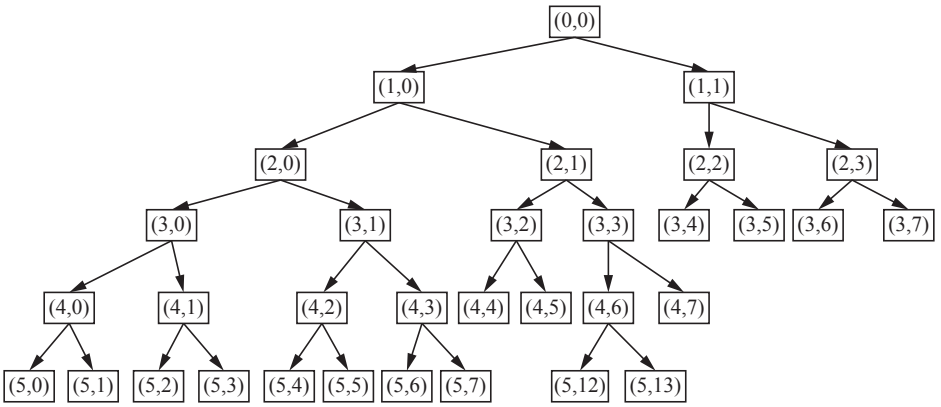


图 2 声音信号的 Bark 尺度小波包分解结构

Fig. 2 Wavelet packet decomposition of sound signal based on Bark scale

$$\boldsymbol{C}=[\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_N]\begin{bmatrix}\lambda_1 & \cdots & 0\\ \vdots & & \vdots\\ 0 & \cdots & \lambda_N\end{bmatrix}\begin{bmatrix}\boldsymbol{u}'_1\\ \vdots\\ \boldsymbol{u}'_N\end{bmatrix}= \lambda_1\boldsymbol{u}_1\boldsymbol{u}'_1+\lambda_2\boldsymbol{u}_2\boldsymbol{u}'_2+\cdots+\lambda_N\boldsymbol{u}_N\boldsymbol{u}'_N \tag{4}$$

式中特征值从大到小递减$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_N$。

3) 前 K 个特征值的确定。特征值λ_i, $i=1,2,\cdots,N$, 代表了特征向量所携带的信息量, 特征值越大说明对应的特征向量所携带的信息量越大。取前 K 个特征值对应的特征向量可以近似地构造出 $\boldsymbol{C}$, 即

$$\boldsymbol{C}\approx\lambda_1\boldsymbol{u}_1\boldsymbol{u}'_1+\lambda_2\boldsymbol{u}_2\boldsymbol{u}'_2+\cdots+\lambda_K\boldsymbol{u}_K\boldsymbol{u}'_K, K\ll N \tag{5}$$

式中 K 值可以通过式 (6) 确定:

$$\eta_K=\sum_{i=1}^{K}\lambda_i/\sum_{j=1}^{N}\lambda_j \tag{6}$$

计算前 K 个特征值之和占全部特征值之和的比重来衡量。

4) 计算频谱投影。选取矩阵 $\boldsymbol{U}$ 中前 K 个成分, 组成特征向量 $\boldsymbol{U}_K=(\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\cdots,\boldsymbol{\mu}_K)$, $\boldsymbol{U}_K\in R^{N\times K}$。计算频谱投影, 即投影矩阵

$$\boldsymbol{X}_K=\boldsymbol{X}\boldsymbol{U}_K \tag{7}$$

$\boldsymbol{X}_K$即为当前小波包分解系数重构的频谱投影特征。

我们对样本声音进行如图 2 所示的 Bark 尺度的小波包分解, 并得到为 17 个分解系数重构的频谱投影, 即 BSP 特征, 将作为 RF 训练与识别的特征。

1.4 随机森林 (RF) 识别

RF 是一种利用多棵决策树分类器来对数据进行判别的集成分类器算法[8], 其输出结果是由决策树输出的类标签的数量而定。这里, 将各个小波包结点分解生成的 BSP 特征结合 RF 分类器, 对动物声音样本进行训练和识别。其过程如图 3 所示, 通过自助重采样技术, 从训练样本第 $i(i=1,2,\cdots,17)$个结点的 BSP 特征集$\boldsymbol{W}_i=\{\boldsymbol{X}_k^1,\boldsymbol{X}_k^2,\cdots,\boldsymbol{X}_k^Q\}$中自助重采样, 生成新的$s$个训练样本集。然后这$s$个训练样本集, 按照决策树的构建方法生长成$s$颗决策树, 并组合在一起形成第$i$个森林。由这 s 棵决策树构造出第i个 RF 与第i个结点的 BSP 特征集相对应。每个 BSP 特征集都要生成一个 RF, 因此一共生成 17 个 RF。

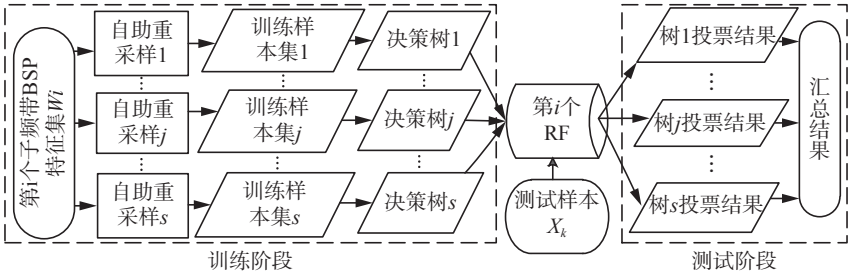


图 3 随机森林 (RF) 训练与识别过程

Fig. 3 The train and recognition process of random forests

RF 对测试样本的识别过程如下。首先, 把测试样本各个 Bark 频率群生成的 BSP 特征$\boldsymbol{X}_k$分别放在相应 RF 的s棵决策树的根节点。根据决策树判别规则向下传递直到决策树的叶子节点。叶子节点对应的类标签就是这棵决策树对特征$\boldsymbol{X}_k$所属类别所做的投票。根据 17 个子频带生成的 RF 中每棵决策树的投票结果, 统计 17 个 RF 中所有投票总和, 其中获得投票数最多的类标签就是测试

样本对应类标签l。

2　声音样本与参数设置

2.1　声音样本集

如表 1 所示, 实验使用的 40 种纯净动物叫声来自 Freesound[16]声音数据库, 分成鸟类和哺乳类; 4 种环境声音, 为录音棒录制的环境背景声音。每种声音有 30 个样本, 实验中随机选取 20 个样本作为训练样本, 其余 10 个样本作为测试样本。对声音文件统一处理, 将其都转换成: 采样率为 8 kHz, 量化精度为 16 bits, 单声道, 且长度为 2 s 左右 wav 格式的声音片段。实验对所有的声音样本归一化处理并采用 Hamming 窗进行分帧。

表 1　声音样本集

Table 1　Sound sample set

声音种类	声音构成
鸟类	1) 翠鸟; 2) 董鸡; 3) 鹤; 4) 黑水鸡; 5) 黄腰太阳鸟; 6) 蓝知更鸟; 7) 公画眉; 8) 水秧鸡; 9) 唐纳雀; 10) 鹈鹕; 11) 燕子; 12) 雨燕; 13) 贼鸥; 14) 绣眼; 15) 小水鸟; 16) 田云雀; 17) 天鹅; 18) 冠纹柳莺; 19) 黄喉地莺; 20) 金丝雀; 21) 海鸥; 22) 八哥; 23) 白面鸡; 24) 斑鸠; 25) 北美夜莺; 26) 北森莺; 27) 捕蝇鸟; 28) 布谷鸟; 29) 苍鹭; 30) 母鹧鸪
哺乳动物	31) 巴塞特猎犬; 32) 草原土拨鼠; 33) 大猩猩; 34) 蝙蝠; 35) 狗; 36) 狐狸; 37) 猎豹; 38) 绵羊; 39) 牛; 40) 山羊
环境声音	流水声; 风声; 公路噪声; 说话噪声

2.2　实验参数设置

1) 帧

在短时傅里叶变换过程中, 每帧帧长为 32 ms, 帧移为帧长一半。

2) 特征

小波包分解采用 db2 基函数, 频谱投影参数 K 通过实验确定。在对比实验中, 声谱图投影特征[17-18]的投影参数 K 取 5; 梅尔频率倒谱系数 (mel frequency cepstrum coefficient, MFCC), 采用 24 阶三角滤波器组, 提取 12 维离散余弦变换系数; 幂归一化倒谱系数 (power normalized cepstrum coefficients, PNCC), 采用 32 阶的 Gammatone 滤波器, 提取 12 维离散余弦变换系数。

3) 随机森林 (RF) 分类器

其主要参数有两个, 一个是决策树中非叶节点分裂时预选特征成分的数量m, 另一个是 RF 中决策树的个数k。综合考虑该文实验样本数量和实验结果, 设定$k=500, m=5$。利用 RF 进行 3 次识别, 然后取均值作为最终结果。

3　实验及结果

3.1　BSP 中 K 的选取

通过纯净声音的 BSP 结合随机森林 (RF) 训练和测试, 确定 BSP 参数 K。在实验中, 我们在没有背景声音的条件下确定 BSP 特征中 K 的选取, K 代表投影矩阵$\boldsymbol{X}_K$中选取的前 K 个特征向量。如图 4 所示, 当$K \leqslant 5$时, 随着 K 的增加, 测试样本的识别率迅速增加, 当$K \geqslant 5$时, 随着 K 的增加, 测试样本的识别率并无明显提升。出于计算代价和性能表现的权衡, 在下面实验中, K 取 5。

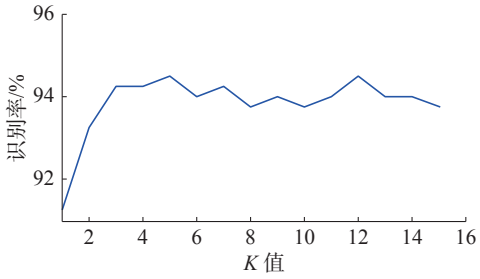


图 4　参数 K 与识别率

Fig. 4　Parameter K and its recognition rate

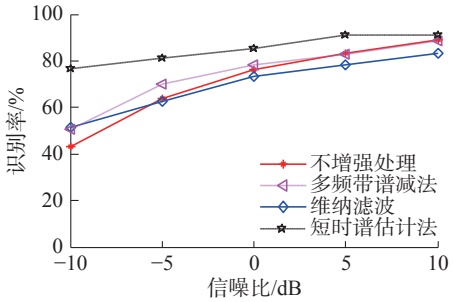
3.2　声音信号增强

使用维纳滤波[19]、多频带谱减法[20]和短时谱估计法[21]对声音进行增强处理, 然后提取 BSP 特征, 分别进行 RF 的识别率测试, 并选出最有效的声音增强算法。

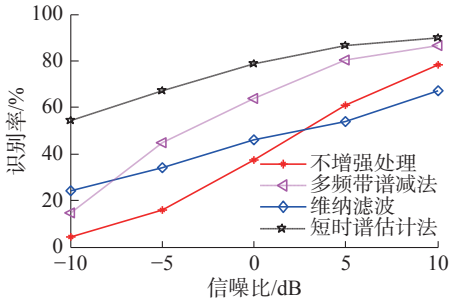
为了减少同一声音事件在不同信噪比及不同噪声环境下, 因增强处理带来信号失真的差异, 实验中我们对纯净的训练声音样本也都分别进行维纳滤波、多频带谱减法和短时谱估计法的增强处理。对测试样本, 在分别添加信噪比为–10 dB、–5 dB、0 dB、5 dB 和 10 dB 的 4 种环境声后, 再进行相应 3 种增强方法处理。在随后的实验中也采取这种方法。

实验结果如图 5 所示。结果表明, 在信噪比为 10 dB 时, BSP 结合 RF 具有 80% 以上的平均识别率。但在不同环境不同信噪比下, 不做声音增强处理的识别率, 整体上低于 3 种声音增强处理的识别率。说明 3 种声音增强算法一定程度上都能消除背景声音的影响。尤其, 在低于 0 dB 的情况下, 特别是低于–5 dB, 3 种声音增强算法消除

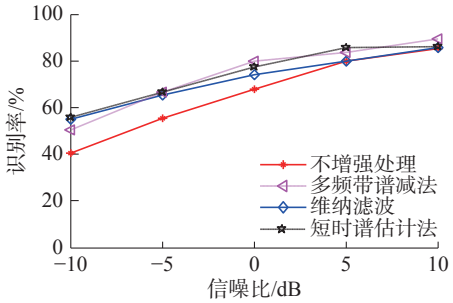
噪声的作用最为明显。从图 5 可以看出，嘈杂说话声环境和流水声环境在−5 dB 以上信噪比时、风声环境在 5 dB 及以上信噪比时，维纳滤波的识别率逐渐低于不增强处理时的识别率。多频带谱减法和短时谱估计法，对 4 种环境各种信噪比下，则都保持较高的识别率。在不同环境声不同信噪比的条件下，短时谱估计法有最佳的识别性能，因此在之后的实验中采用短时谱估计增强方法对声音信号进行加强。



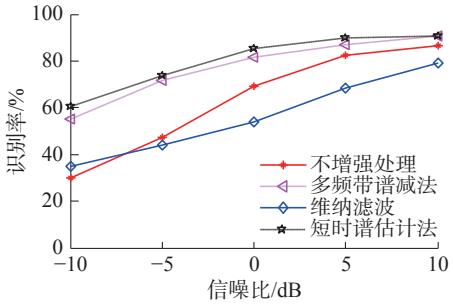
(a) 流水噪声环境



(b) 风声噪声环境



(c) 公路噪声环境



(d) 说话噪声环境

图 5　4 种不同噪声环境下不同增强处理方法的识别效果

Fig. 5　Results of different enhancement process in four kinds of noisy environments

3.3　BSP 与常用特征比较

把 BSP 特征与 SPD[12]、声谱图投影[17-18]、PNCC 和 MFCC 等 4 种常用特征进行 RF 的训练与识别实验比较。

首先，无噪声条件下的实验，结果如表 2 所示。BSP、SPD、声谱图投影、PNCC 和 MFCC 等 5 种特征对动物声音的识别率都达到 90% 以上，其中，目前对声音事件识别最有效的 SPD，识别效果略好于该文的 BSP。

表 2　无噪声条件下不同方法的比较

Table 2　Comparing different method in non-noise condition　%

方法	识别率
BSP	94.5
SPD	96.3
声谱图投影特征	94.3
PNCC	93.5
MFCC	91.6

其次，在不同噪声环境不同信噪比条件下的 5 种特征的平均识别率实验结果如表 3 所示。利用流水声、风声、公路声和嘈杂说话声，模拟真实复杂环境噪声。取信噪比−10 dB、−5 dB、0 dB、5 dB、10 dB 和 15 dB，分别与 4 种噪声环境进行混合，用于 RF 训练并测试 5 种不同特征提取的平均识别率。不同噪声环境下的平均识别率如表 3 所示。从表 3 中可以看到，在不同环境不同信噪比条件下，BSP 的平均识别率达到 80.5%，比 SPD、声谱图投影、PNCC 和 MFCC 等 4 种特征分别高出 11.4%、9.6%、17.1% 和 50.5%。

表 3　在不同噪声环境下的平均识别率

Table 3　Average accuracy in different noisy environments　%

噪声类型	不同特征提取方法的平均识别率				
	BSP	SPD	声谱图投影特征	PNCC	MFCC
流水	85.8	74.4	77.8	73.3	28.1
风声	77.7	66.2	64.0	57.0	33.3
公路	76.7	62.6	65.1	48.8	24.9
说话	81.9	73.3	76.5	74.4	33.6
平均	80.5	69.1	70.9	63.4	30.0

由于 MFCC 在低信噪比的识别率明显低于其他 4 种特征，随后，我们只比较其他 4 种特征的识别效果。图 6 表示 BSP、SPD、声谱图投影和 PNCC 等 4 种特征，在 4 种噪声环境下，信噪比为 −10 dB、−5 dB、0 dB 和 5 dB 时的识别率。从图中可以看出，在信噪比小于 0 dB 时，BSP 特征的识别率明显高于其他 3 种特征。

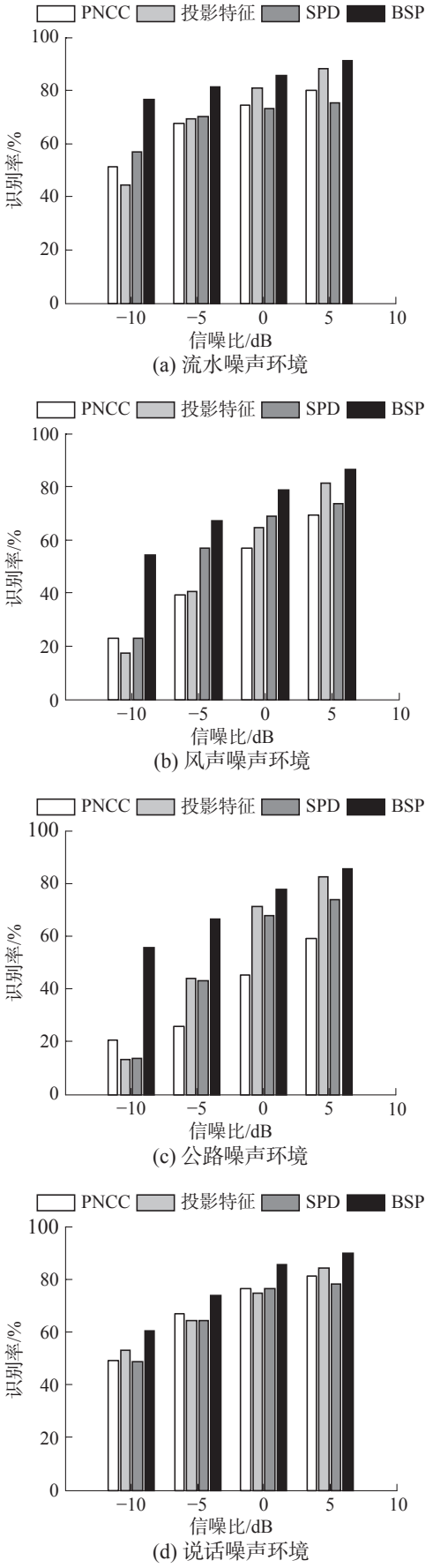


图 6　4 种噪声环境下不同信噪比的识别率

Fig. 6　Recognition rate of four kinds of features in four kinds of noisy environments

3.4　与现有方法及分类器的比较

把该文提出 BSP-RF 与 MP-SVM[10]、PC-SVM[11]和 SPD-KNN[12]等声音事件检测识别的 3 种方法进行比较, 结果如表 4 所示。从表 4 中可以看出, 本文方法 BSP-RF 在低信噪比情况下的识别率, 与文献[10-12]中的方法相比有较大提高。BSP-RF 在–10 dB 的情况下, 依然能够保持平均 60% 以上的识别率, 效果尤为明显。其次, 我们进行 BSP 结合 SVM, BSP 结合 KNN 的实验。结果表明, 对于 BSP 特征而言, 采用 RF 对各种环境下不同信噪比动物声音的识别效果优于 SVM 与 KNN。

表 4　6 种方法的平均识别率

Table 4　Average recognition rate of six kinds of methods　%

方法	纯净	20 dB	10 dB	0 dB	–10 dB	平均
BSP-RF	94.5	91.4	89.5	81.8	61.8	83.8
MP-SVM[10]	86.3	80.7	56.5	29.5	14.6	53.5
PC-SVM[11]	91.4	88.8	87.5	78.6	42.2	77.7
SPD-KNN[12]	97.3	94.6	94.3	78.2	45.3	81.9
BSP-SVM	87.3	85.1	81.8	71.6	51.5	75.5
BSP-KNN	86.8	84.2	77.2	64.0	40.9	70.6

4　讨论

4.1　RF、SVM 与 KNN 对 BSP 识别性能的分析

从表 4 可以看出, RF 的平均识别率高于 KNN 和 SVM。特别是在–10 dB 的情况下, 分别比 KNN 和 SVM 高出 20.9%、10.3%, 说明 RF 比 KNN、SVM 更适用于 BSP 特征的分类识别。由于 BSP 特征把声音信号分解成 17 个频带, 每个频带只包含部分的声音信息, 用这些不完整的信息进行 KNN 分类会造成识别率的下降。KNN 是基于距离的分类方法, 某个特征维度之间的差异值过大可能很大程度上影响其他特征维度, 同时 KNN 不能给出决策树那样的分类规则, 所以 BSP 特征在 KNN 的分类效果低于 RF。SVM 适用于高维度、分类数目少、小样本的分类识别, BSP 的特征维度相对较小且实验中包含 40 类动物声音, 所以文中方法不适合采用 SVM 进行分类。RF 是基于决策树的分类规则挖掘不同特征维度之间的关系, 同时结合不同频带之间投票结果, 可以提高 BSP 特征的分类精度, 所以 RF 比 KNN 以及 SVM 更适用于 BSP 特征的分类。

4.2　环境声音对动物声音的影响

为了分析环境声音对动物声音在各个 Bark 频率群的影响, 我们给出纯净的翠鸟声音和加入信噪比为–10 dB 背景声音后各个 Bark 频率群的能量分布。从图 7 中可以看出各个 Bark 频率群

的能量的变化以及背景声音对翠鸟声音在各个 Bark 频率群的影响。

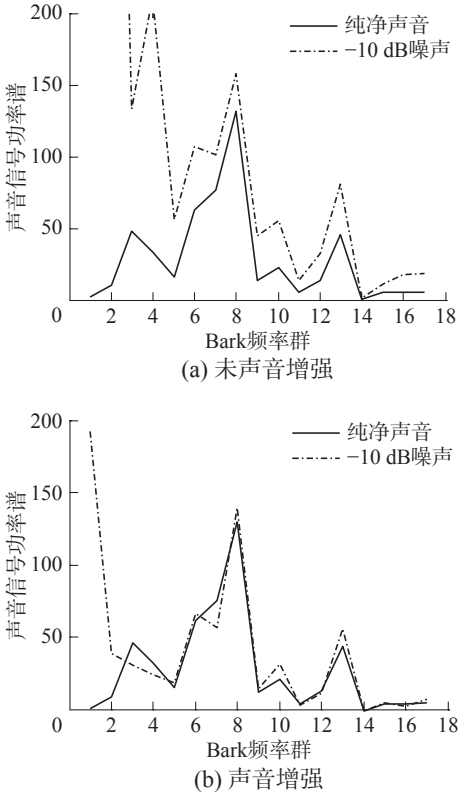


图 7 翠鸟的各个 Bark 频率群的能量分布

Fig. 7 The energy distribution of kingfisher in each Bark-frequency group

图 7(a) 和 (b) 是翠鸟声音经过声音增强前后的各个 Bark 频率群的能量分布。从图中可以看出背景声音对 Bark 频率群 1 ~ 4, 即低频部分的影响比较大, 对于高频部分的影响相对比较小。经过短时谱估计法声音增强后, 可以消除大部分背景声音的影响, 但影响依然存在。该文结合经过 Bark 尺度小波包结构, 把声音信号分解成 17 个投影特征。这样, 可以有效地平衡背景声音对部分 Bark 频率群的影响, 有利于识别率的提高。

4.3 动物及环境声音与重构频谱投影

1) Bark 尺度的小波包分解的本质

Bark 尺度的小波包分解的本质, 就是把声音信号按人类听觉敏感程度, 对声音信号进行频带划分, 再进行不同尺度的小波分析。动物声音, 即便在各种环境中, 受到不同信噪比的环境声音的干扰, 只要人类听觉能感知到, 就意味着它存在不同于环境声音的 Bark 频率群。而本文提出的 Bark 尺度的小波包分解系数重构频谱投影, 就是分离出这些相关频率群频谱的关键成分。这些 Bark 频率群的频谱, 必然为每一种动物声音的特色或独有。用这些频谱的投影, 进行随机森林 (RF) 的投票, 必定是高分。而与那些与背影声音同频率群的成分, 虽然在投票中难获高分, 但多个频率群共同投票后, 仍然能保持较高的得分优势。

2) 错误检测的分析

表 5 给出加入−5 dB 风声后, 16 类容易出现错误检测的情况 (另外 24 类基本上能够正确识别, 限于空间, 表 5 中未列出)。从表 5 可以看出, 在−5 dB 风声下, 第 10 类的测试样本全部被错误检测, 其中有 9 个测试样本错分到第 19 类中; 第 24、28、38、39 这 4 类测试样本也都全部被错误检测成第 19 类。同时, 发现大部分被错误检测的样本, 都被检测成第 19 类。

表 5 加入−5 dB 风声噪声测试样本错分情况

Table 5 Wrong test samples' condition in −5 dB wind noise

错分类 标签	4	5	10	19	20	23	24	28	29	30	34	35	37	38	39	40
4	5	1														
5		6														
10			0	9												
19				10												
20					4						1					
23						4				6						
24				10			0									
28				10				0								
29				6					1	1		1				
30										6		1				
34				6							4					
35				8								1				
37				9									1			
38				10										0		
39				10											0	
40				4			1	2						1		1

观察图 8(a) 10 类 (鹈鹕)、(b) 19 类 (黄喉地莺) 和 (c) 38 类 (绵羊) 声音分别在−5dB 风声下的声谱图, 可以发现它们的相似之处。其中, 低频部分, 即 0~800 Hz 部分相似度较高; 3 张声谱图在 0.5 s 之后, 高低频部分都很相近。也就是说, 这 3 张频谱图, 高低频部分有 80% 左右是相近的。这就可能造成大部分 Bark 频率群频谱投影的相近或相等, 从而造成了测试样本的错误检测。

从实验结果进一步观察到, 在加入风声噪声的情况下, 大部分错分的样本被错分到第 19 类; 在加入嘈杂说话声时则大部分的样本被错分到第 2 类; 在加入公路噪声时则大部分的样本被错分到第 39 类。这说明测试样本错分的原因和加入噪声的类型有关。

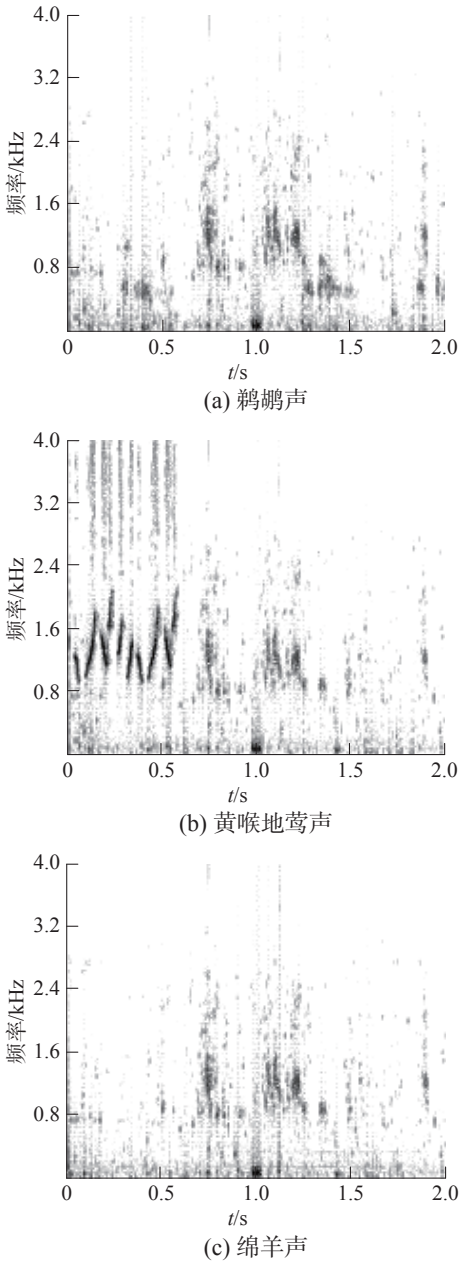


图 8　3 种不同动物在-5 dB 风声下的频谱

Fig. 8　Spectrums of three kinds of animal in -5 dB wind noise

但是, 它们作为不同的动物的声音, 自然环境下, 能被人类听觉感知到, 必然有区别于环境声音的成分存在, 即有不同于背景声音的 Bark 频率群存在。因此, 根据该文方法的原理, 这种差别可以通过小波包分解结构及随机森林投票策略的适当调整来识别。进而, 本文提出的方法可以在各种背景声音中, 识别各种不同信噪比的动物声音。

3) 更深层次的识别

对于非平稳的环境及动物声音, 如在特定的背景声音环境下, 各种动物声音混在一起, 时强时弱等情况, 有可能影响 RF 投票结果。对于这种情况, 我们可以考虑帧一级的 RF 投票。如, 声音信号按 32 ms 分帧, 只要动物声音不是在 32 ms 内同时发生, 我们依然可以通过 RF 投票确定每一帧可能的动物声音, 并进一步来判断出可能的多种的动物声。这种情况下, 这种方法甚至可以识别出人类很难识别的非平稳及混合的各种动物声音。

5　结论

实验表明, 在–10 dB 以上信噪比环境下, 在未对声音信号进行增强处理的情况下, 该文提出的方法对于动物声音识别有较好的效果。而短时谱估计声音增强结合 BSP 特征与随机森林的方法, 不论是低信噪比还是高信噪比声音环境, 对各种环境中的动物声音检测都有较好的效果。

提出的方法能胜任于自然环境下各种低信噪比动物声音识别的原因如下: 1) 采用短时谱估计声音增强算法, 一定程度上抑制了环境声音的影响。2) Bark 尺度的小波包分解是基于人耳基底膜的工作原理, 环境声音对于不同 Bark 频率群的影响是不一样的, 因此结合各个 Bark 频率群的特征信息作为决策依据, 一定程度上能够提高识别率。3) 采用多随机森林决策的方法有效地消除了环境声音对部分 Bark 频率群特征的影响。

在后续的工作中, 将结合深度学习相关方法, 围绕如何在多个声音重叠的情况下实现各个声音事件的检测与识别做进一步的研究。

参考文献:

[1] MITROVIC D, ZEPPELZAUER M, BREITENEDER C. Discrimination and retrieval of animal sounds[C]//Proceedings of the 12th International Multi-Media Modelling Conference Proceedings. Beijing, China: IEEE, 2006: 339–343

[2] JANČOVIC P, KÖKÜER M, ZAKERI M, et al. Bird species recognition using HMM-based unsupervised modelling of individual syllables with incorporated duration modelling[C]//Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 559–563.

[3] RAGHURAM M A, CHAVAN N R, BELUR R, et al. Bird classification based on their sound patterns[J]. International journal of speech technology, 2016, 19(4): 791–804.

[4] BARDELI R. Similarity search in animal sound databases [J]. IEEE transactions on multimedia, 2009, 11(1): 68–76.

[5] POTAMITIS I, NTALAMPIRAS S, JAHN O, et al. Automatic bird sound detection in long real-field recordings: applications and tools[J]. Applied acoustics, 2014, 80: 1–9.

[6] ZHANG Xiaoxia, LI Ying. Adaptive energy detection for bird sound detection in complex environments[J]. Neurocomputing, 2015, 155: 108–116.

[7] 魏静明, 李应. 利用抗噪纹理特征的快速鸟鸣声识别[J]. 电子学报, 2015, 43(1): 185–190.
WEI Jingming, LI Ying. Rapid bird sound recognition using anti-noise texture features[J]. Acta electronica sinica, 2015, 43(1): 185–190.

[8] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1): 5–32.

[9] FENG Zuren, ZHOU Qing, ZHANG Jun, et al. A target guided subband filter for acoustic event detection in noisy environments using wavelet packets[J]. IEEE/ACM transactions on audio, speech, and language processing, 2015, 23(2): 361–372.

[10] WANG Jiacheng, LIN Changhong, CHEN Bowei, et al. Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation[J]. IEEE transactions on automation science and engineering, 2014, 11(2): 607–613.

[11] DENNIS J, TRAN H D, LI Haizhou. Spectrogram image feature for sound event classification in mismatched conditions[J]. IEEE signal processing letters, 2011, 18(2): 130–133.

[12] DENNIS J, TRAN H D, CHNG E S. Image feature representation of the subband power distribution for robust sound event classification[J]. IEEE transactions on audio, speech, and language processing, 2013, 21(2): 367–377.

[13] LI Ying, WU Zhibin. Animal sound recognition based on double feature of spectrogram in real environment[C]// Proceedings of 2015 IEEE International Conference on Wireless Communications and Signal Processing. Nanjing, China: IEEE, 2015: 1–5.

[14] LAINE A, FAN J. Texture classification by wavelet packet signatures[J]. IEEE Transactions on pattern analysis and machine intelligence, 1993, 15(11): 1186–1191.

[15] KARMAKAR A, KUMAR A, PATNEY R K. Design of optimal wavelet packet trees based on auditory perception criterion[J]. IEEE signal processing letters, 2007, 14(4): 240–243.

[16] Universitat Pompeu Fabra. Repository of sound under the creative commons license, Freesound.org[DB/OL]. [2018-03-13]. http://www.freesound.org.

[17] KIM H G, MOREAU N, SIKORA T. Audio classification based on mpeg-7 spectral basis representations[J]. IEEE transactions on circuits and systems for video technology, 2004, 14(5): 716–725.

[18] DENG Shiwen, HAN Jiqing, ZHANG Chaozhu, et al. Robust minimum statistics project coefficients feature for acoustic environment recognition[C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2015: 8232–8236.

[19] CHANG Kangming, LIU S H. Gaussian noise filtering from ECG by Wiener filter and ensemble empirical mode decomposition[J]. Journal of signal processing systems, 2011, 64(2): 249–264.

[20] PALIWAL K, WÓJCICKI K, SCHWERIN B. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain[J]. Speech communication, 2010, 52(5): 450–475.

[21] 刘翔, 高勇. 一种引入延迟的语音增强算法[J]. 现代电子技术, 2011, 34(5): 85–88.
LIU Xiang, GAO Yong. Speech enhancement algorithm with leading-in delay[J]. Modern electronics technique, 2011, 34(5): 85–88.

作者简介：

黄鸿铿，男，1993 年生，硕士研究生，主要研究方向为声音事件检测、信息安全。

李应，男，1964 年生，教授，博士，主要研究方向为多媒体数据检索、声音事件检测、信息安全。获授权发明专利 10 项。发表学术论文 20 余篇。