基于多随机森林的低信噪比声音事件检测

李 应[1,2],印佳丽[1,2]
(1. 福州大学数学与计算机科学学院,福建福州 350116;2. 网络系统信息安全福建省高校重点实验室,福建福州 350116)

摘 要: 论文针对各种背景声音中低信噪比声音事件的检测问题,提出把背景声音与声音事件混合,形成带噪声样本来训练分类器. 在预处理阶段,使用基于经验模态分解与 2-6 级固有模态函数的投票方法,对背景声音与声音事件端点进行预测并估算信噪比. 接着使用子带能量分布方法,提取声音数据的特征. 最后,论文将背景声音与声音事件样本库中所有声音样本按照估算的信噪比相混合,生成混合声音特征训练多随机森林,用于低信噪比声音事件的检测. 实验证实,所提出的方法可以用于各种声场景下低信噪比声音事件的检测,并能在信噪比为 -5dB 的情况下保持 67.1% 的平均检测率.

关键词: 声音事件检测; 信噪比; 经验模态分解; 子带能量分布; 随机森林
中图分类号: TP391.42 **文献标识码:** A **文章编号:** 0372-2112 (2018)11-2705-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.11.018

Sound Event Detection at Low SNR Based on Multi-random Forests

LI Ying[1,2], YIN Jia-li[1,2]
(1. *College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian* 350116, *China*;
2. *Key Lab of Information Security of Network Systems (Fuzhou University), Fuzhou, Fujian* 350116, *China*)

Abstract: For sound event detection under various background noises at low SNR, this paper proposes a method that mixes the background noises with sound events into noisy samples to train classifiers. In the pre-processing stage, we use a voting method based on 2th to 6th intrinsic mode functions (IMFs) that generated from empirical mode decomposition (EMD), to detect the endpoint of sound events and estimate the SNR. Then subband power distribution (SPD) is used to extract features from audio data. Finally, we mix the background noise and all the sound event samples in the sound event database according to the estimated SNR, and then extract the noisy samples features to train multi-random forests (M-RF) for the detection of the sound events in low SNR environment. The experiment proves that the proposed method has the ability to recognize sound events in various acoustic scenes at low SNR, and can remain an average accuracy rate of 67.1% at -5dB.

Key words: sound event detection; signal-to-noise ratio (SNR); empirical mode decomposition; subband power distribution; random forests

1 引言

低信噪比声音事件检测,就是试图检测、分类和识别嵌入在各种噪声和混响音频信号中的相对微弱的声音对象. 近来,声音事件检测引起广泛关注. 它不只是由于随着网络中多媒体数据的快速增长,基于音频数据的多媒体搜索具有极大的应用价值,同时,声音事件检测也是分析环境的关键组成之一.

关于声音事件检测,目前的研究包括:吵闹环境下特定目标声音事件检测方法[1];声音事件的特征[2~4]及声音事件的分类器[5,6];背景/前景检测、声音事件分类和声音事件定位方法[7];声音场景、室内声音事件以及室内综合声音事件的检测与分类方法[8,9];特定环境下特定声音的检测方法[10]等.

这些方法对于声音事件的检测都取得一定的效果. 然而,特征提取过程都有不同程度地对声音事件的特征即待测声音信号的特征本身的结构造成影响. 虽然用于特征缺失的谱掩饰估算算法能有效去除被场景

收稿日期:2016-12-02;修回日期:2018-06-03;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61075022);福建省自然科学基金(No.2018J01793)

声音干扰的声音事件的特征[11],但也屏蔽了声音事件的部分特征. 而在白噪声的情况下,短时估计特征缺失的方法[12],容易出现声音事件的关键特征的缺失,影响检测效果. 谱减法[13]对不同频段的信号都进行不同处理,仍然不可避免地对声音事件中的部分属性造成影响. 虽然多频带谱减法[14]对谱减法做出了改进,但相关问题依然存在.

为了避免在抑制场景声音的同时,对声音事件信号结构的影响,从而在低信噪比下得到了更高的检测率,本文提出用背景声音与声音事件混合的声音特征来训练分类器. 在分类器模型的训练过程中,背景声音按不同信噪比与声音事件进行混合,得到声音事件在各种声场景下的声音数据,对分类器进行训练. 在检测处理中,通过希尔伯特-黄变换(Hilbert-Huang transform, HHT)中的经验模态分解(empirical mode decomposition, EMD)[15]预测声音事件和背景声音的边界点. 根据预测出的声音事件和背景声音的边界点,估算声音事件的信噪比. 根据估算的信噪比,把预测出的背景声音与声音事件样本进行混合,用混合声音事件样本集的特征训练检测声音事件的分类器.

对于检测声音事件的分类器,参考相关文献[16~18]和已有工作[19,20],本文基于随机森林(random forests, RF),提出用多随机森林(multi-random forests, M-RF)检测低信噪比声音事件的方法. 对于各种声音事件及其背景声音的信号特征,采用声音信号子带能量分布[11](subband power distribution, SPD)的统计特征.

2 声音事件检测架构

图1是本文提出的基于随机森林的低信噪比声音事件检测的基本架构[20]. 以图1为基础,图2是用于低信噪比声音事件检测的多随机森林架构.

如图1所示,测试中把待测声音信号 $y(t)$ 通过经验模态分解(EMD),预测出背景声音部分 $n(t)$ 和声音事件部分 $s(t)$. 通过背景声音 $n(t)$ 和声音事件 $s(t)$,估算待测声音事件的信噪比 l_s. 把背景声音 $n(t)$ 根据信噪比 l_s,与声音事件样本库中的 M 种声音事件进行混合,提取混合声音数据的子带能量分布(SPD)的统计特征,生成特征集 $\boldsymbol{W}^s$,用 $\boldsymbol{W}^s$ 训练随机森林 RF_s. 用随机森林 RF_s 对声音事件 $s(t)$ 的特征 $\boldsymbol{w}^e$ 进行检测.

一般情况下,对声音信号中声音事件的信噪比的估算存在偏差. 实验表明,尤其在低信噪比时,如果对信噪比的估算出现偏差,训练出的随机森林可能无法对声音事件进行准确检测. 因此,我们把图1的随机森林架构进一步扩展成图2所示的多随机森林架构.

在多随机森林架构中,对于声音信号,我们同时用实际估算的信噪比 l_s 值及其相近的两个信噪比 l_{sh} 和 l_{sl} ($l_{sh} > l_s > l_{sl}$),分别与声音事件样本混合成三组声音集. 用三组混合声音集分别训练三个 RF 分类器 RF_{sh}, RF_s, RF_{sl}. 在对声音事件进行检测时,分别用 RF_{sh}, RF_s, RF_{sl} 对声音事件进行检测,最后通过三个随机森林中的所有决策树投票确定待测声音的检测结果.

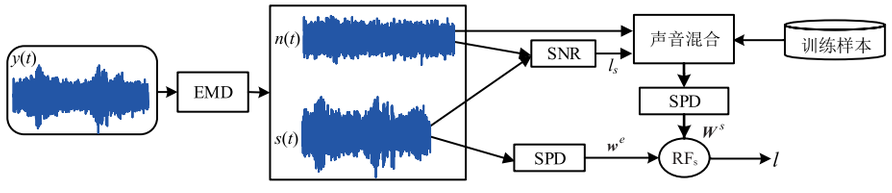


图1 背景声中检测声音事件的随机森林架构

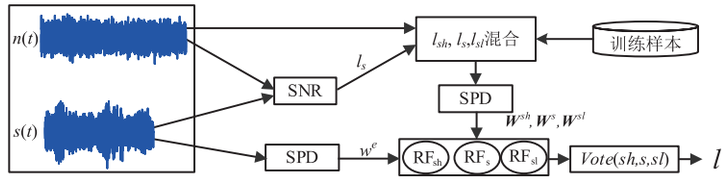


图2 背景声中检测声音事件的多随机森林架构

3 低信噪比声音事件检测

3.1 经验模态分解与信噪比估算

经验模态分解能依据信号自身的特性将原始信号 $y(t)$ 自适地分为 n 级固有模态函数 $L_i(t)$ 的线性叠加[15],即,

$$y(t) = \sum_{i=1}^{n} L_i(t) + r_i(t) \tag{1}$$

其中,$r_i(t)$ 为残余函数.

由于声音信号的主要信息集中在低频部分,1级固有模态函数 $L_1(t)$ 以声音信号中的高频成分为主,对于声音事件端点检测的贡献有限. 因此,本文选取 $i=2$,

3,…,6 的 $L_i(t)$,用于声音事件端点的估计,并将得到的 5 种不同端点估计的结果,经投票确定为最终的端点预测结果.

根据这种方法,以火烈鸟叫声声音事件为例,图 3 中蓝色部分为声音信号波形图,红色部分为端点预测结果. 其中,高位表示包含声音事件 $s(t)$,低位表示仅包含背景声音 $n(t)$. 图 3(a)、(b)、(c)、(d)、(e)、(f)为火烈鸟声音事件及其在各种声场景下 0dB 声音事件的端点预测结果. 通过图 3 可以看出,该方法在 0dB 下能够基本预测出声音事件段.

将声音信号 $y(t)$ 预测为声音事件段 $s(t)$ 与背景声音段 $n(t)$ 之后,我们对声音事件的信噪比 l_s 进行的估计.

$$l_s = 10\log_{10}\frac{\sum_{t=1}^{M}[s(t))]^2 - \ell\sum_{t=1}^{N}[n(t)]^2}{\ell\sum_{t=1}^{N}[n(t)]^2} \tag{2}$$

其中,M 为声音事件片段 $s(t)$ 的长度,N 为背景声片段 $n(t)$ 的长度,$\ell = M/N$. 由于分离后的声音事件段中含有背景声音成分,对声音事件段的能量值产生影响. 又因为在大部分情况下,噪音能量在短时间内不变,因此使用 $\ell\sum_{t=1}^{N}[n(t)]^2$ 作为声音事件段中噪音能量的估计,可以降低背景声音对能量值的影响.

3.2　声音信号的子带能量分布特征

频率子带能量分布(SPD)[11]通过对每一个频率子带中不同等级能量的概率密度统计,将频谱图转换为频率子带能量分布. 通过提取子带能量分布的统计特征,即子带能量分布特征,用于声音事件的检测.

以一个长度 2 秒的火烈鸟叫声音事件为例,子带能量分布计算过程如图 4 所示. 包括:(1)将声音信号转化成如图 4(a)所示的 Gammatone 频谱图 $S(f,t)$[21];(2)将 $S(f,t)$ 转化为图 4(b)所示归一化的频谱图 $G(f,t)$;(3)统计归一化频谱图子带上的能量分布,生成图 4(c)所示子带能量分布 $H_R(f,b)$;(4)对 $H_R(f,b)$ 做增强处理,生成图 4(d)所示增强子带能量分布 $H(f,b)$.

提取声音信号增强子带能量分布特征:

(1)将增强子带能量分布 $H(f,b)$ 进行量化和映射,转化成三种单色子图 $M_c(f,b)$:

$$M_c(f,b) = \begin{cases} \dfrac{l_2 - l_1}{H(f,b) - l_1}, & l_1 < H(f,b) < l_2 \\ 1, & l_2 \le H(f,b) \le u_1 \\ \dfrac{u_2 - u_1}{u_2 - H(f,b)}, & u_1 < H(f,b) < u_2 \\ 0, & \text{other} \end{cases} \tag{3}$$

其中 $c = \{red, green, blue\}$,red,green 和 blue 子图的映射参数 $\{l_1, l_2, u_1, u_2\}$ 分别为:red $\{0.375, 0.625, 0.875, 1.125\}$,green $\{0.125, 0.375, 0.625, 0.875\}$ 和 blue $\{-0.125, 0.125, 0.375, 0.625\}$.

(2)如图 5 所示,将每张子图等分为 $D \times D$ 个局部子块 $\boldsymbol{L}_{i,j}^c$. 统计子块的信息:

$$w_{i,j}^c = \{\boldsymbol{\mu}(\boldsymbol{L}_{i,j}^c), \boldsymbol{\sigma}^2(\boldsymbol{L}_{i,j}^c)\} \tag{4}$$

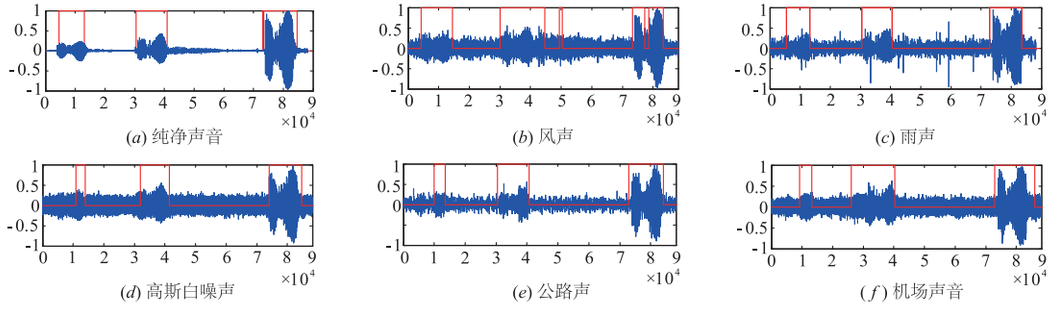
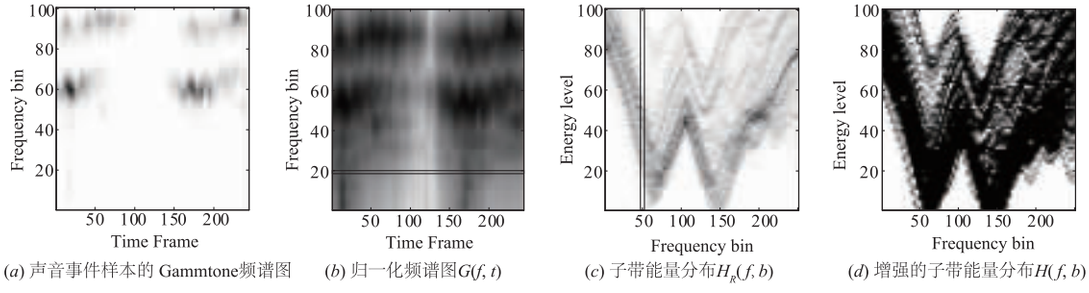


图3　纯净声音及0dB的不同背景声音下声音信号端点预测结果



(a) 声音事件样本的 Gammtone频谱图　(b) 归一化频谱图$G(f, t)$　(c) 子带能量分布$H_R(f, b)$　(d) 增强的子带能量分布$H(f, b)$

图4　SPD图生成过程

其中，$\mu(\boldsymbol{L}_{i,j}^{c})$表示子块中像素值的平均值，$\sigma(\boldsymbol{L}_{i,j}^{c})$则表示子块中像素值的标准差. 子带能量分布特征成分可以表示为

$$w_{i,j}=\{w_{i,j}^{red},w_{i,j}^{green},w_{i,j}^{blue}\} \tag{5}$$

其中，$i,j=1,\cdots,D$. 这样，一个声音信号，可以获取相应的子带能量分布特征 w：

$$w=\{\mu^{red},\sigma^{red},\mu^{green},\sigma^{green},\mu^{blue},\sigma^{blue}\} \tag{6}$$

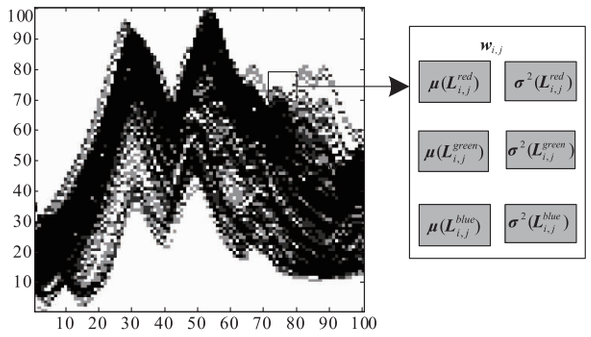


图5 提取SPD图特征$\boldsymbol{w}$

3.3 对子带能量分布特征的重组

本文涉及的特征值 $\boldsymbol{w}$ 包括图 1 中的 $\boldsymbol{W}^{s}$，图 2 中的 $\boldsymbol{W}^{sh}$，$\boldsymbol{W}^{s}$，$\boldsymbol{W}^{sl}$和待测声音事件的特征值 $\boldsymbol{w}^{e}$.

如图 1 所示，我们把端点预测得到背景声音 $n(t)$ 与声音事件样本集中的所有声音事件样本，按照信噪比 l_s进行混合，并提取混合声音事件的子带能量分布特征集 W，用于随机森林分类器的训练.

对训练样本的特征集

$$W=\{w^{1},w^{2},\cdots,w^{i},\cdots,w^{Q}\},$$

$$w^{i}=\{\mu^{i\text{-}red},\mu^{i\text{-}green},\mu^{i\text{-}blue},\sigma^{i\text{-}red},\sigma^{i\text{-}green},\sigma^{i\text{-}blue}\}$$

进行调整. 使得，

$$W=\{\boldsymbol{M}^{red},\boldsymbol{\Sigma}^{red},\boldsymbol{M}^{green},\boldsymbol{\Sigma}^{green},\boldsymbol{M}^{blue},\boldsymbol{\Sigma}^{blue}\},$$

其中，

$$\boldsymbol{M}^{c}=\{\mu^{1c},\mu^{2c},\cdots,\mu^{ic},\cdots,\mu^{Qc}\},$$

$$\boldsymbol{\Sigma}^{c}=\{\sigma^{1c},\sigma^{2c},\cdots,\sigma^{ic},\cdots,\sigma^{Qc}\},$$

$$c=\{red,green,blue\}.$$

3.4 多随机森林检测

在随机森林训练中，分别用 $\boldsymbol{M}^{red}$，$\boldsymbol{\Sigma}^{red}$，$\boldsymbol{M}^{green}$，$\boldsymbol{\Sigma}^{green}$，$\boldsymbol{M}^{blue}$，$\boldsymbol{\Sigma}^{blue}$等 6 个特征子集，训练如图 6 所示的 6 个随机子森林. 对于待测声音事件，我们也将特征 $\boldsymbol{w}^{e}$拆分为 $\boldsymbol{\mu}^{e\text{-}red}$，$\boldsymbol{\sigma}^{e\text{-}red}$，$\boldsymbol{\mu}^{e\text{-}green}$，$\boldsymbol{\sigma}^{e\text{-}green}$，$\boldsymbol{\mu}^{e\text{-}blue}$，$\boldsymbol{\sigma}^{e\text{-}blue}$. 它们也分别输入到相应的随机森林进行测试. 检测结果通过 6 个随机子森林投票得出. 即，如果每个子随机森林中有 500 棵决策树，那么共有 6 * 500 = 3000 棵决策树；这样，每棵决策树投票给一个类，得票最多的类，即为多随机森林的分类结果. 对于 M 棵决策树，n 个样本，m 个特征，其复杂度为 $O(M(mn\log_n))$. 采用 6 个子随机森林，每个子随机森林只根据一种特征做分类，因此随机森林在训练过程中无需做特征选择，复杂度为 $O(Mn\log_n)$.

由于 3.1(2)估算得到的信噪比 l_s可能存在偏差，为此，通过对估算信噪比 l_s的分布分析，增设两个信噪比，即 l_{sh}和 l_{sl}. 实验中，我们把测试样本中所有估计信噪比高于 l_s的样本的估计信噪比平均值作为 l_{sh}. 低于 l_s的样本估计信噪比平均值为 l_{sl}.

根据 l_{sh}、l_s和 l_{sl}三个不同的信噪比，我们把背景声音 $n(t)$分别与声音事件样本混合成三组声音集，并分别提取三组混合声音事件的子带能量分布特征集 $\boldsymbol{W}^{sh}$，$\boldsymbol{W}^{s}$和 $\boldsymbol{W}^{sl}$. 采用图 6 的 6 个随机子森林的方式，训练图 2 所示的架构中的多随机森林 RF_{sh}、RF_{s}和 RF_{sl}，并用于实现低信噪比声音事件的检测.

4 实验

4.1 实验设置

为确保实验的可靠性，我们采用的声音事件样本集由 40 种动物声音与 6 种背景声音组成，具体类别及编号如表 1 所示. 40 种纯净动物声音来自 Freesound 声音数据库[22]，每种动物声音有 30 个样本，共 1200 个样本. 每种声音样本中随机选取 20 个样本作为训练样本，其余 10 个样本作为测试样本. 5 种常见背景声音分别为公路声、刮风声、下雨声、流水声以及机场噪声. 这五类噪声都是非平稳噪声，对声音事件有较大的干扰性. 背景声音是以 44.1kHz 的采样频率，分别在相应的背景中录制. 此外，为加入平稳噪声，我们在背景声音集中加入高斯白噪声. 为规范以上声音文件的编码格式和长度，我们将它们统一转换成采样频率为 8kHz、采样精度为 16bit，长度为 2s 的单声道 WAV 格式声音片段.

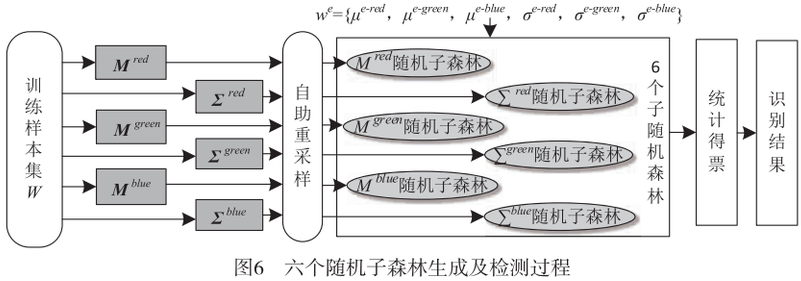


图6 六个随机子森林生成及检测过程

表 1　声音样本类别

声音类别	具体子类别及标签				
声音事件	1. 赭红尾鸲	2. 秃鹰	3. 乌鸦	4. 布谷	5. 鸽子
	6. 鸭子	7. 喜鹊	8. 猫头鹰	9. 知更鸟	10. 海鸥
	11. 天鹅	12. 海燕	13. 雨燕	14. 海狮	15. 羊
	16. 老虎	17. 鲸鱼	18. 牛羚	19. 狼	20. 蝙蝠
	21. 熊	22. 猫	23. 黑猩猩	24. 鹿	25. 狗
	26. 海豚	27. 驴	28. 大象	29. 马	30. 狮子
	31. 猴子	32. 猪	33. 海豹	34. 牛	35. 火烈鸟
	36. 马蹄声	37. 骆驼	38. 草地鹨	39. 潜鸟	40. 花栗鼠
背景声音	平稳噪声	高斯白噪声			
	非平稳噪声	雨声,机场噪声,公路噪声,流水噪声,风声			

实验中,我们将在 -5dB,0dB,5dB 三种信噪比的 6 种背景声音中比较声音事件检测率. 每次按照一种信噪比将一种背景声音与纯净测试样本混合,生成各种信噪比的样本,作为输入声音信号. 从而,实验将得到 3 种不同信噪比下 6 种背景声音的检测率.

实验中,随机森林分类器中的决策树数量 DT_number = 500,节点特征数量 Feature_number = 5. 随机子森林中的每个子森林,也同样取决策树数量 DT_number = 500 和节点特征数量 Feature_number = 5.

SVM 直接利用 LIBSVM[23] 工具箱进行 SVM 的训练和测试建模,其中,设定核函数为径向基核函数,惩罚因子 $c = 2$,核参数 $g = 2.8$.

4.2　多随机森林与随机森林的比较

针对信噪比估计偏差,我们挑选出测试样本中所有估计信噪比高于 l_s 的测试样本,求得信噪比平均值 l_{sh},以及信噪比低于 l_s 的样本信噪比平均值 l_{sl},把背景声音 $n(t)$ 与声音事件样本 $s(t)$ 分别根据 l_{sh} 和 l_{sl} 混合,并生成另外两组混合声音事件的子带能量分布(SPD)特征集 $\boldsymbol{W}^{sh}$ 和 $\boldsymbol{W}^{sl}$. 与图 1 中 $\boldsymbol{W}^{s}$ 一起,如图 2 所示,用 $\boldsymbol{W}^{sh}$,$\boldsymbol{W}^{s}$ 和 $\boldsymbol{W}^{sl}$ 训练多随机森林(M-RF) RF_{sh}、RF_s 和 RF_{sl},并进行各种信噪比下的声音事件的检测实验.

实验结果如图 7 所示. 我们可以看到图 2 所示的 M-RF 架构在低信噪比下能够提升检测率. 图 7 中,-5dB 下 M-RF 的检测率 67.1%,而 RF_s、RF_{sh} 和 RF_{sl} 则分别为 57.5%,55% 和 61%.

4.3　与现有方法的比较

M-RF 方法与现有的四种方法[9,11,24,25] 在 -5dB、0dB、5dB 三种信噪比及六种背景声音环境下的声音事件平均检测率如表 2 所示. 结果显示,本文的 M-RF 方法明显优于其他方法. 在 -5dB,0dB,5dB 三种信噪比中,M-RF 平均检测率为 86.8%,比 SPD + KNN 方法高出近 6%,比 SIF 方法高出 12.7%. 而 MFCC + SVM 方法在 -5dB,0dB,5dB 三种信噪比中的平均检测率只有 33.0%,MP-feature 则为 26.1%. 从表 2 可以看出,当信噪比为 0dB 时,M-RF、SPD + KNN、SIF 方法均能保持 80% 以上的检测率. 但当信噪比为负时,所有方法的检测率呈快速下降,但本文方法依然保持 67.1% 的检测率.

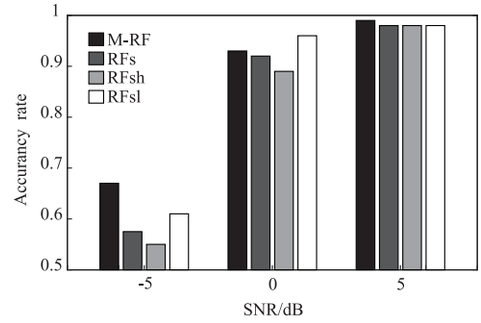


图7　多随机森林与随机森林对低信噪比声音事件的检测率

表 2　不同方法在三种信噪比及六种背景声音环境下的平均声音事件检测率(%)

方法	不同信噪比下检测率			
	5dB	0dB	-5dB	平均
M-RF(本文)	99.1	94.2	67.1	86.8
MP-feature[9]	36.2	23.6	18.5	26.1
SPD + KNN[11]	94.3	88.0	60.2	80.8
SIF[24]	85.3	81.0	56.2	74.1
MFCC + SVM[25]	45.2	30.2	23.6	33.0

5　讨论

5.1　M-RF 低信噪比声音事件检测结果分析

图 8 为 -5dB 的六种背景声下,用本文的 M-RF 方法的检测结果. 图中坐标为 (x,y) 的点的颜色,对应属于 y 类的样本被检测为 x 类的个数. 在实验测试过程

中,每一类动物声音包含 10 个测试样本,40 类共 400 个样本. 从背景声音来看,在高斯白噪声下的声音事件检测率最高. 高斯白噪声是平稳噪声,功率谱密度服从均匀分布,因此在信号频谱图上,声音事件与背景声叠加后高频部分仍然是原信号的高频部分,SPD 所反映出来的高频特征依然有较好的代表性.

我们进一步对图 8 所示的六种背景声音下的检测结果进行统计. 按照表 1 中的各种动物叫声事件的顺序,把编号 1-13 的动物声音事件及六种背景声音下 -5dB 的平均检测结果归结为表 3. 与图 8 相对应,表 3 中的行表示需要检测的声音类别,列则表示检测结果. 当信噪比为 -5dB 时,秃鹰、布谷、鸽子和天鹅四种鸟类声音的平均检测率在 90% 以上. 其中,鸽子的检测率达到了 100%. 但是,雨燕叫声和海鸥叫声分别只有 3.3% 和 25% 的检测率,鸭子和喜鹊的叫声分别只有 35% 和 38.3% 的检测率. 48.4% 的雨燕声音样本被分入了海燕叫声,68.3% 的海鸥叫声样本被分到了海燕. 同样,65% 的鸭子叫声、56.7% 的喜鹊叫声被分到了海燕. 我们从表 3 可以看出,在信噪比低至 -5dB 的情况下,第 1 类、第 6 类、第 7 类、第 10 类和第 13 类声音的大部分测试样本都被分类为第 12 类.

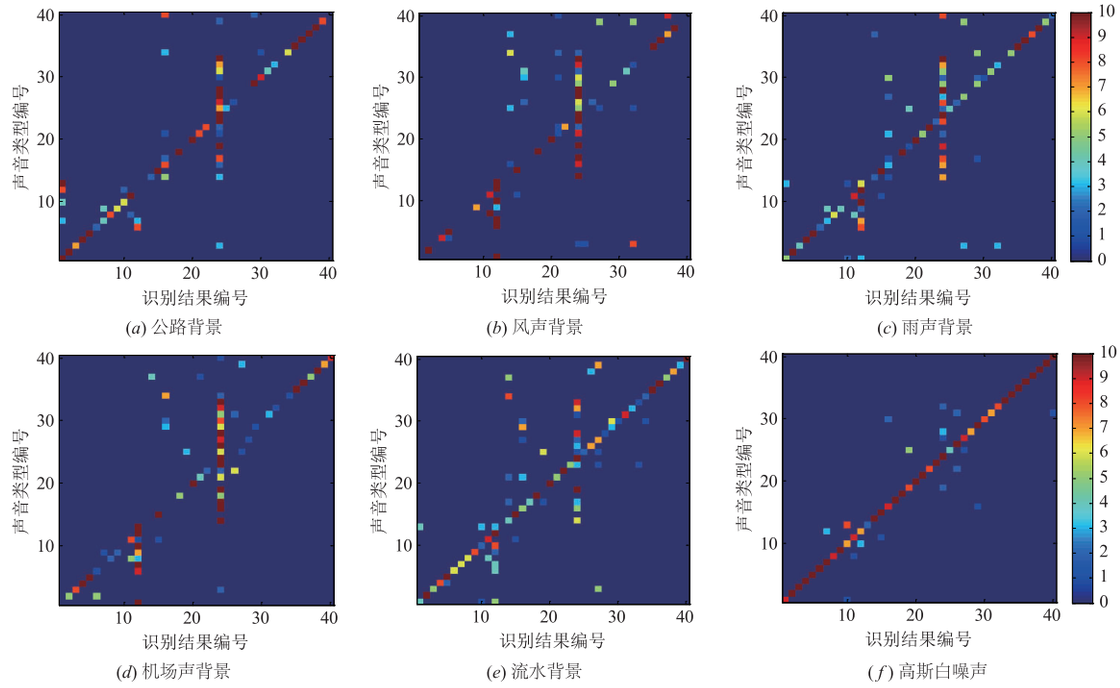
同样,我们可以从图 8 中看出,在信噪比低至 -5dB 的情况下,在各种背景声音下,第 14、16、17、19、23、26、28、32、33 类声音的大部分测试样本都被错分到第 24 类中.

由图 8 及表 3 可知,同样在 -5dB 的低信噪比环境下,有的声音事件保持了较高的检测率,有的则无法正确检测.

5.2 低信噪比声音事件特征分析

在自然环境中,声音事件一般有它的独特之处,即它与背景声音有不同的一面. 图 9(*a*)和(*b*)分别是第 7 类喜鹊声音的 Gammatone 频谱及 SPD 图. 图 9(*c*)和(*d*)是第 12 类海燕声音的 Gammatone 频谱及 SPD 图. 可以看出,这两类声音的频率及能量等级分布相似. 喜鹊叫声信号的高能量部分在 20 到 50 频带及 100 到 140 频带之间. 而图 9(*c*)中海燕叫声的高能量等级部分包含了喜鹊叫声的高能量等级的频率区域. 不同的是,喜鹊在 0 -20 频带能量等级高于海燕,130 -150 频带高能部分比例分布比海燕略多. 反映在 SPD 图上,如图 9(*b*)(*d*)所示,在 0 -20 频带能量等级高于海燕,在能量等级 70 -80、频带 130 -150 的区域比海燕的图 9(*d*) 颜色深.

同样图 9(*e*)和(*f*)是第 33 类海豹声音的 Gammatone 频谱及 SPD 图. 图 9(*g*)和(*h*)是第 24 类鹿鸣声音的 Gammatone 频谱及 SPD 图. 同样地,两类声音在频率及能量等级分布相似. 其中,高能量部分大致相同. 但是,图 9(*f*)海豹声能量等级 70 -95、频带在 60 -100 的区域与鹿鸣声的图 9(*e*) 显示出不同. 因此,如果能根据这些微弱的不同,进行进一步的分析,本文方法可以作为异常声音事件的检测与监控的有效手段.



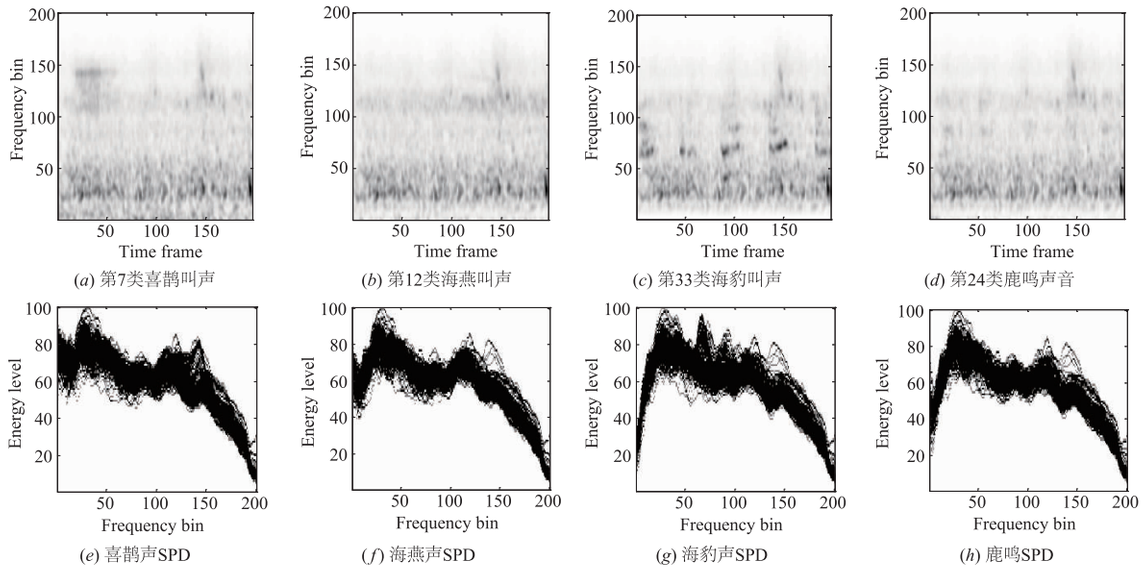
(*a*) 公路背景 (*b*) 风声背景 (*c*) 雨声背景

(*d*) 机场声背景 (*e*) 流水背景 (*f*) 高斯白噪声

图8 -5dB下各种背景下的分类结果

表3　鸟类叫声在6种 −5*dB* 背景噪声下平均检测结果(%)

	1 赭红尾鸲	2 秃鹰	3 乌鸦	4 布谷	5 鸽子	6 鸭子	7 喜鹊	8 猫头鹰	9 知更鸟	10 海鸥	11 天鹅	12 海燕	13 雨燕	其他
1 赭红尾鸲	46.7									6.6		46.7		
2 秃鹰		91.7				8.3								
3 乌鸦			56.7											43.3
4 布谷				95	5									
5 鸽子					100									
6 鸭子						35						65		
7 喜鹊	5						38.3					56.7		
8 猫头鹰								51.7			43.3	5		
9 知更鸟							15		61.7			23.3		
10 海鸥	6.7									25		68.3		
11 天鹅											90	1.7		8.3
12 海燕	13.3						5			3.4		78.3		
13 雨燕	28.3									20		48.4	3.3	



(*a*) 第7类喜鹊叫声　(*b*) 第12类海燕叫声　(*c*) 第33类海豹叫声　(*d*) 第24类鹿鸣声音

(*e*) 喜鹊声SPD　(*f*) 海燕声SPD　(*g*) 海豹声SPD　(*h*) 鹿鸣SPD

图9　−5dB风声背景下声音信号Gammatone频谱及SPD

5.3　异常的处理

如 −5dB 风声背景下的喜鹊叫声和 −5dB 机场背景下的雨燕叫声事件,在正常环境下是不可能出现的;出现可能意味着异常. 实验中,我们通过模拟方式,在声音事件的检测内容中加入低信噪比异常声音事件. 由于对于某些低信噪比及异常声音事件,其在频带及能量分布上可能与背景声音相差极其微弱,因此增加检测的难度.

对于可能出现的低信噪比异常声音事件,需要进一步的分析. 如通过建立经验相关模型,可以针对可能的异常声音事件进行细节分析之后,再用本文的方法进行检测.

在实际应用中,对于某种场景而言,可能发生的声音事件有限. 因此,声音事件样本库中的声音事件数量也有限,按照图1的 RF_s 或图2的M-RF架构,把样本库中与场景相关的声音事件进行混合,并建立 RF_s 或 $RF_{sh}-RF_s-RF_{sl}$,可以在线进行. 这样使得在各种声场景下检测低信噪比声音事件可以在线进行.

对于待测声音事件信噪比估算的偏差,引起检测率降低. 考虑到背景声音的非平稳性,分离出的环境声音与其它时间段的环境声音存在偏差. 解决的方法可以选择多段代表性的非平稳场景声音,分别与样本库

中的声音事件进行混合,生成多个 RF,最后结果由多个 RF 的检测结果,进一步投票确定.

6 结论

本文提出一种能够在各种声场景下、有效提高低信噪比下检测率的声音事件检测方法. 该方法把待测声音事件中的场景声音,与声音事件样本库相结合,通过 SPD 图提取声音数据的特征,生成判别待测声音事件的多随机森林. 利用这种方法生成的随机森林,可以在特定场景中,实现低信噪比下,声音事件的检测. 实验结果表明,该方法能使声音事件与场景声音信噪比为 -5dB 的情况,保持平均精度 67% 以上声音事件的检测率. 与 MP、MFCC、SIF 和 SPD + KNN 等方法相比,一定程度上说,我们所提出的这种方法能解决低信噪比情况下,声音事件的检测问题.

参考文献

[1] Zuren Feng, Qing Zhou, Jun Zhang, et al. A target guided subband filter for acoustic event detection in noisy environments using wavelet packets[J]. IEEE Trans on Audio, Speech, and Language Processing, 2015, 23(2): 361 - 372.

[2] Grzeszick R, Plinge A, Fink G A. Bag-of-features methods for acoustic event detection and classification[J]. IEEE/ACM Trans on Audio, Speech, and Language Processing, 2017, 25(6): 1242 - 1252.

[3] Ren Jian feng, Jiang Xu dong, Yuan Jun song, et al. Sound-event classification using robust texture features for robot hearing[J]. IEEE Trans on Multimedia, 2017, 19(3): 447 - 458.

[4] Ye Jia xing, Kobayashi T, Murakawa M. Urban sound event classification based on local and global features aggregation [J]. Applied Acoustics, 2017, 117: 246 - 256.

[5] Ozer I, Ozer Z, Findik O. Noise robust sound event classification with convolutional neural network[J]. Neurocomputing, 2018, 272: 505 - 512.

[6] 李艳雄,王琴,张雪,等. 基于凝聚信息瓶颈的音频事件聚类方法,电子学报,2017,45(5):1064 - 1011.
LI Yan xiong, Wang Qin, Zhang Xue, et al. Audio events clustering based on agglomerative information bottleneck [J]. Acta Electronica Sinica, 2017, 45(5): 1064 - 1011. (in Chinese)

[7] Phan H, Maab M, Mazur R, et al. Random regression forests for acoustic event detection and classification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 20 - 31.

[8] Stowell D, Giannoulis D, Benetos E, et al. Detection and classification of acoustic scenes and events[J]. IEEE Trans on Multimedia, 2015, 17(10): 1733 - 1746.

[9] Wang J, Lin C, Chen B, et al. Gabor-based nonuniform scale-frequency map for environmental sound classification in home automation[J]. IEEE Trans on Automation Science and Engineering, 2014, 11(2): 607 - 613.

[10] Sharma A, Kaul S. Two-stage supervised learning-based method to detect screams and cries in urban environments [J]. IEEE Trans on Audio, Speech, and Language Processing, 2016, 24(2): 290 - 299.

[11] Dennis J, Ttran H D, Chng E S. Image feature representation of the subband power distribution for robust sound event classification[J]. IEEE Trans on Audio, Speech, and Language Processing, 2013, 21(2): 367 - 377.

[12] Seltzer M, Raj B, Stern R. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition[J]. Speech Communication, 2004, 43(4): 379 - 393.

[13] Yamashita K, Shimamura T. Nonstationary noise estimation using low-frequency regions for spectral subtraction [J]. IEEE Signal Processing Letters, 2005, 12(6): 465 - 468.

[14] Sunil K, Philipos L. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise [A]. IEEE international Conference on Acoustics, Speech and Signal Processing[C]. Orlando, USA: IEEE, 2002. 13 - 17.

[15] Huang H, Pan J. Q. Speech pitch determination based on hilbert-huang transform[J]. Signal Processing, 2006, 86(4): 792 - 803.

[16] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 - 32.

[17] Pang H, Lin A, Holford M, et al. Pathway analysisusing random forests classification and regression[J]. Bioinformatics, 2006, 22(16): 2028 - 2036.

[18] Unella K L, Hayward L B, Scgal J, et al. Screening large-scale association study data: exploiting interactions using random forests[J]. BMC Genetics, 2004, 11(5): 32 - 37.

[19] Wei Jingming, Li Ying. Specific environmental sounds recognition using time-frequency texture features and random forest[A]. International Congress on Image and Signal Processing [C]. Hangzhou, China, 2013. 1705 - 1709.

[20] Lin Wei, Li Ying. Lower SNR sound event recognition using noisy training sample[A], International Congress on Image and Signal Processing[C]. Shenyang, China, 2015. 1448 - 1453.

[21] Pour A F, Asgari M, Hasanabadi M R. Gammatonegram based speaker identification[A]. International E-Conference on Computer and Knowledge Engineering[C]. Hong Kong, China, 2014. 52 - 55.

[22] Universitat P F. Repository of sound under the creative commons license, freesound. org [DB/OL]. http://www. freesound. org,2012 - 5 - 14.
[23] Chang C C,Lin C J. LIBSVM:a library for support vector machines[J]. ACM Trans on Intelligent Systems and Technology,2011,2(3):27.
[24] Dennis J,Tran H D,Li H. Spectrogram image feature for sound event classification in mismatched conditions[J]. IEEE Signal Processing Letters,2011,18(2):130 - 133.
[25] Zheng F,Zhang G,Song Z. Comparison of different implementations of MFCC[J]. Journal of Computer Science and Technology,2001,16(6):582 - 589.

作者简介

李　应　男,1964 年出生,福建闽清人,福州大学数学与计算机科学教授,主要研究领域为信息安全,多媒体数据检索.
E-mail:fj_liying@ fzu. edu. cn

印佳丽　女,1993 年出生,江苏盐城人,硕士,主要研究领域为模式识、环境声音检测.