SRT项目报告

计科*30* 李志远
November 13, 2014

1 绪论

在这次的SRT实践中，我前期学习了一些乐理及信号处理的基本方法，主要的参考书籍为[1]。我主要负责了音频识别与生成方面的一些工作，具体包括了音频起音点检测(Audio Onset Detection) 和音频音调检测(Audio Key Detection)。下面将对此简述。

2 音频起音点检测(audio onset detection)

这部分算法的主要思路[2]是先把音频划分成一个个小片段，对每个片段做瞬时傅里叶变换(STFT), 把各频率的强度分量取模相加得到函数$f(m)$。

定义

$$f_{onset}(m) := \frac{f(m) - f(m-1)}{f(m)} \tag{1}$$

我们认为f_{onset}的极值点就对应了起音点。

2.1 具体步骤

1. 对输入的离散时域信号$s(n)$做瞬时傅里叶变换，

$$S_k(m) := \sum_{n=mh}^{mh+N-1} w(n-mh)s(n)e^{-j\Omega_N k(n-mh)} \tag{2}$$

其中N是每次取样窗长，$\Omega_N = \frac{2\pi}{N}$, h为取样间隔，$k = 0, 1, 2 \ldots N-1$代表了频率,$w(n)$是Hanning Window。

2. 定义

$$f(m) := \sum_{l=1}^{L} |S_l(m)| \tag{3}$$

其中$L = \frac{N}{2} + 1$。此处因为对称性我们只取前一半的频率。

3. 如同(1),我们定义了f_{onset} [1] ,通过求f_{onset}的极值来寻找起音点。

2.2 样例图像

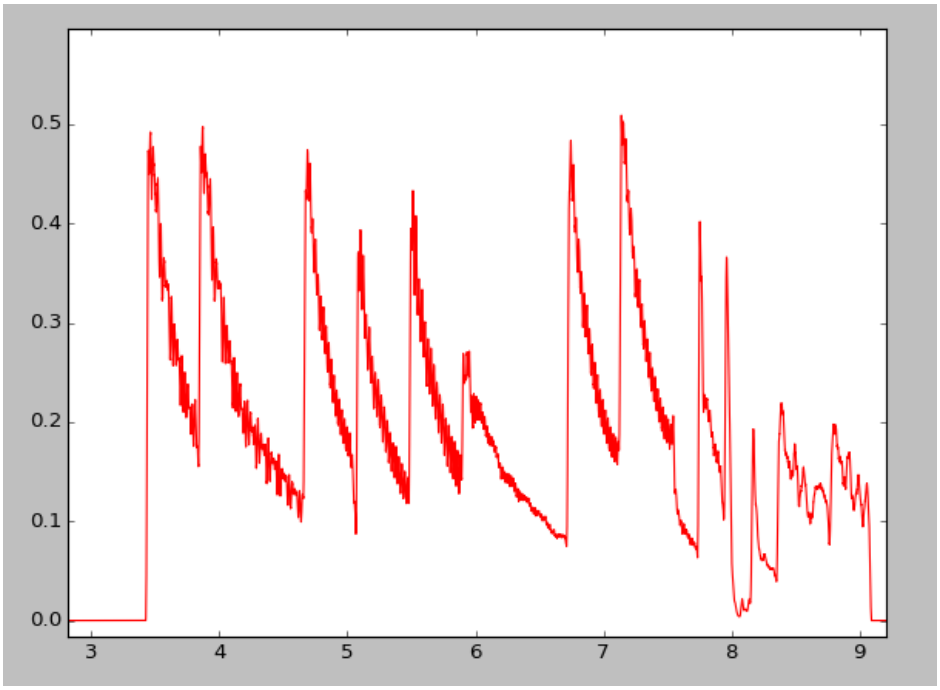

Figure 1: 对样例音频的识别原始数据包络

图1为原始数据的包络，图2为f和f_{onset}的图像，其中蓝色曲线为f的图像，红色曲线为f_{onset}的图像，可以看到，原始数据的包络和f的图像大致相同，而且f_{onset}的图像的极大值点与原音频的起音点吻合得很好。(原始波形数据已做过正规化处理）

2.3 算法改进

实践中，在[2]中提出的算法遇到了很大的困难：对于不同性质的乐器，其音质有所不同，如果单单是取最大值，则乐器产生的泛音也有可能被当成起音点检测到造成误判。

[1] $f_{onset} \approx (\log f(m))'$, 因为人耳对响度的感受是与强度的对数成正比，所以f_{onset}的极值也就反应了听觉感受到的各频率分量响度增强最快的那个时间点

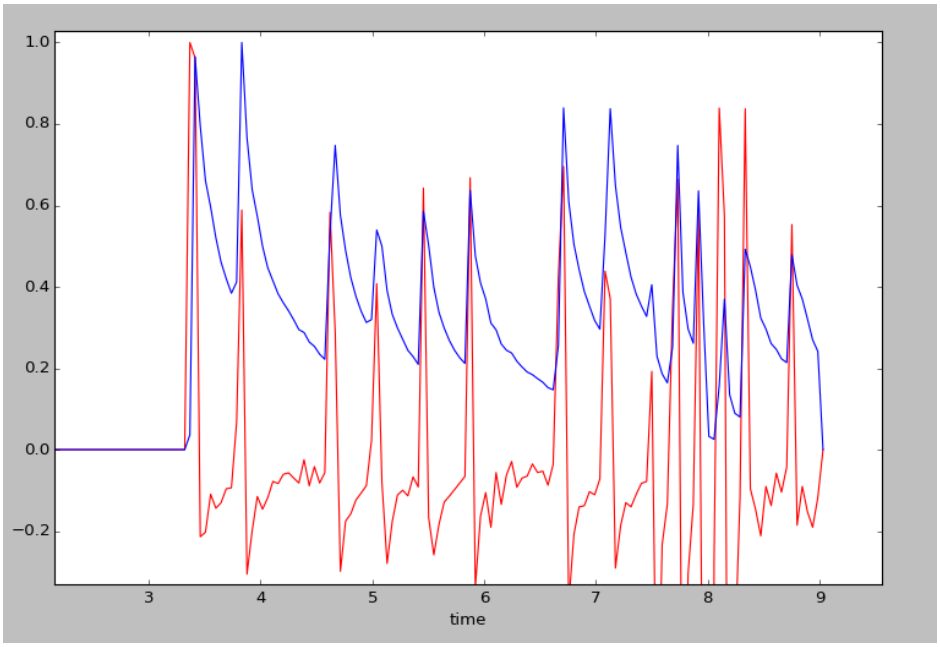

Figure 2: 对样例音频的识别f,f_{onset}

为了解决这个问题，我对起音点的判定添加了一些附加条件。首先是f_{onset}自身要大于某一个阈值，*threshhold1*。但是这样虽然避免了一些泛音的问题，却也引起了一些漏音的现象，尤其是当音频平均音量与峰值音量相差不多的时候，正规化处理后$f(m)$一直处于较高的水平，从而使$f_{onset}(m)$值较小。所以我又修改了判断条件，f_{onset}可以有一个更低的阈值,*threshhold2*，但是这个时候对f的局部最大值也有了相应的要求，要求$max(f[m-1:m+2])$大于*threshhold3*。

```
seq_onset = []
for m in range(M):
    if (f_onset[m]>f_onset[m-1]) and (f_onset[m]>f_onset[m+1]) and\
    ((f_onset[m]>threshhold1) or ((f_onset[m]>threshhold2) and (max(f[m-1:m+2])>threshhold3))):
        seq_onset += [m]
```

Figure 3: 阈值判断函数

最后经过不断地调整与尝试，得到threshhold1=0.5, threshhold2=0.4, threshhold3=0.6，这是一组效果比较理想的参数，显著地提升了起音点识别的准确性。

3 音频音调检测(Audio Key Detection)

通过对识别出的起音点判断做CQT(Constant-Q Transform)[3], 我们获得了82个feature，分别代表了从$C\#1$ 到$A\#7$这82个不同频率上的强度。利用开源的SVM软件包libsvm，对2201个不同乐器不同频率的单音wav文件[2] （实际样本数为3002）进行了1 vs all 的监督式学习，得到了一个可靠的分类器。[3]

详细原理参见[2]。

3.1 SVM训练参数的调节

在linear kernel的one vs all的SVM算法中，需要调节的参数只有c。c越大，意味着SVM得到的模型bias越大，variance越小。

通过"0.03 0.1 0.3 1 3 10 30" 这种近似等比的调节方式，得到了最优的参数为$c = 3$。

3.2 结果及正确率

在数据集上的正确率为97%, 按照80-20划分的Cross Validation 为89%。

对来源为实际生活的数据的测试：用ipad上的Piano HD应用弹奏”小星星“，前十四个音符均被准确识别，返回了正确的时间和音高。

References

1. Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons, 2012.
2. Giovanni Costantini, Massimiliano Todisco, Renzo Perfetti, Roberto Basili, and Daniele Casali. Svm based transcription system with short-term memory oriented to polyphonic piano music. In *MELECON 2010-2010 15th IEEE Mediterranean Electrotechnical Conference*, pages 196–201. IEEE, 2010.

[2]数据来源：http://theremin.music.uiowa.edu/MISPost2012Intro.html

[3]本来是$C1$ 到$B7$ 这84个feature,但是在取平均，降低向量维数的时候$C1$和$B7$产生了边缘效应，可靠度降低，故舍弃。

3. Benjamin Blankertz. The constant-q transform.