AI 吉他评测算法技术报告

1. 评测算法总述

吉他测评算法核心是让机器通过确定性算法对演奏给出合理的评价。吉他演奏的评价方法较为多样，为了便于量化研究，我们选取音准、流畅度作为评价演奏质量的关键指标，对评测系统进行建模。记音准为 a，流畅度为 f，则用户所得分数 s 表示为两者的加权均值：

$$s = \frac{ap_1 + fp_2}{2} \quad (0 \le a, f, s, p_1, p_2 \le 1)$$

其中p_1、p_2分别为音准与流畅度所占权重。

模型假设用户演奏与标准演奏越接近，则所得分数越高。具体地，如果用户错弹或者漏弹，则 a 值应相应降低；如果用户节奏不连贯，则 f 值也应相应降低。权值p_1、p_2根据测评的难易要求确定。

综上，算法归结于如何根据用户演奏计算 a 值、f 值。形式化描述为，寻找一组合理的模型$(\tilde{a}, \tilde{f})$，使得：

$$\begin{cases} a = \tilde{a}(w, s) \\ f = \tilde{f}(w, s) \end{cases}$$

其中 w 为用户演奏，s 为标准演奏，吉他演奏最终都以离散音频信号的形式表示和参与运算。由于音乐与机器之间存在一定的语义鸿沟[1]，必须建立合适的算法提取音乐信息。

2. 音准评分关键技术

人类听觉之所以能识别一段音频中所包含乐音的音高，是因为各种乐音具有不同的基音频率（Fundamental Frequency）。对于机器而言，利用离散傅里叶变换(DFT)将时域的音频信号转换到频域，便可方便地在频谱中筛选出基音频率，从而识别音高。

但问题在于，除基频外，在乐器频谱中还包括多次谐波，设基音频率为f_b，则 k 次谐波频率为kf_b（谐波直接决定了乐器的音色）。此外，实际频谱中不可避免地混有噪声分量。这些因素导致一个关键问题：算法该选择哪个频率分量作为乐音的基频。

一种常见的基频检测方法为谐波积谱法（Harmonic Product Spectrum, HPS）[2]。设乐音信号 x(n)的频谱为$X_n(e^{j\omega})$。由于谐波频率为kf_b，若对频谱进行内积得到P_n：

$$P_n\left(e^{j\omega}\right) = \prod_{k=1}^{N} X_n(e^{jk\omega})$$

则由于 k 次谐波角频率$k\omega$与基频$\omega = 2\pi f_b$保持固定的 k 倍关系，不论原始频谱中基频分量是否具有峰值，$P_n\left(e^{j\omega}\right)$一定会在基频处出现峰值。这便有效降低了基频落在无关频率上的概率。

谐波积谱法的缺陷在于，它将频谱上给定频率范围内的所有分量都考虑了进去，包括噪声分量，导致噪声分量可能影响识别结果。另一种改进算法[3]只关心音阶内各音符对应的基音频率，除非噪声频率恰好落在音阶的基音频率上，否则噪声不会影响识别结果。

根据 12 平均律，若中央 C 频率定调为$f_c = 130.81\,Hz$，则该音阶内的 12 个半音的基音

频率为

$$f(n) = 2^{\frac{n}{12}} \cdot f_c \ , n = 0,1, \dots ,11$$

改进算法仅考虑 12 个半音的基音频率在频谱中的幅度，而不像谐波积谱法那样对整个频谱进行计算。

具体地，在 DFT 变换得到的幅度谱X(k), k = 0,1, ...中，设频率 f 对应的幅度为$X\left(k^{(f)}\right)$，则$k^{(f)}$满足：

$$k^{(f)} = \left\lfloor \frac{f}{f_s/N} + 0.5 \right\rfloor$$

其中 N 为采样个数，f_s为采样频率

已知 12 平均律中各半音的基音频率$f = f(n)$，便可通过上式得出这些音符在频谱中的幅度$X\left(k^{(f)}\right)$，通过比较各个半音幅度的大小，便可确定源信号为某个半音的可能性，称半音的能量$C(n)$。半音的能量与幅度的关系为$C(n) = g(X\left(k'^{(n)}\right))$，其中算法$g$·与$k'$·将在后续章节详细介绍。12 个半音的能量组成了一维向量$\boldsymbol{C} = (c_0, c_1, \dots, c_{11})$，称之为色谱图(Chromagram)。色谱图已经确定了音频到音符的关系，可用于音准分析。

上述算法在分析单音乐器时已经充分，但对于吉他等复音乐器而言，其发音并非仅由一系列单音构成，在同一时刻很可能存在多个单音共同奏响的情况，乐理上称之为和弦，显然算法还需进一步对和弦进行识别。

综上，本节引出了音准评分的两个关键技术：色谱图(Chromagram)以及和弦识别，下面分别详述。

3. 计算色谱图（Chromagram）

已知音频信号 x(n)，对其进行离散傅里叶变换(DFT)，得到幅度谱 X(k)，则色谱图可计算为：

$$C(n) = \sum_{\phi=1}^{2} \sum_{h=1}^{2} \left(\max_{k_0{}^{(n,\phi,h)} \le k \le k_1{}^{(n,\phi,h)}} X(k) \right) \frac{1}{h}$$

其中 n 为一个 8 度内的半音，$n = 0,1, \dots 11$；ϕ表示计算几个 8 度；h 为待计算的谐波数。对于半音 n，算法在幅度谱中从k_0到k_1的范围内寻找峰值。

寻找范围定义为：

$$k_0{}^{(n,\phi,h)} = k_c{}^{(n,\phi,h)} - 2h$$
$$k_1{}^{(n,\phi,h)} = k_c{}^{(n,\phi,h)} + 2h$$

其中

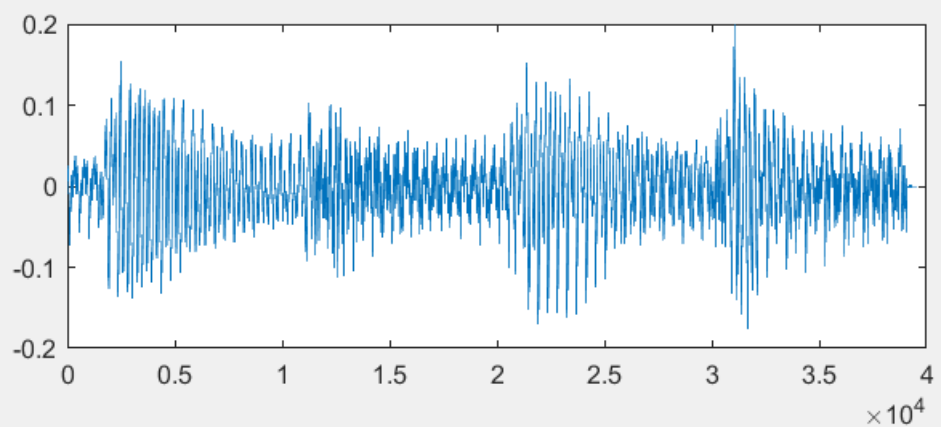
$$k_c{}^{(n,\phi,h)} = k^{(f(n)\cdot\phi\cdot h)} = \left\lfloor \frac{f(n) \cdot \phi \cdot h}{f_s/N} + 0.5 \right\rfloor$$

表示寻找范围的中心。$f(n)$为大字组 8 度音阶内半音 n 对应的基音频率。

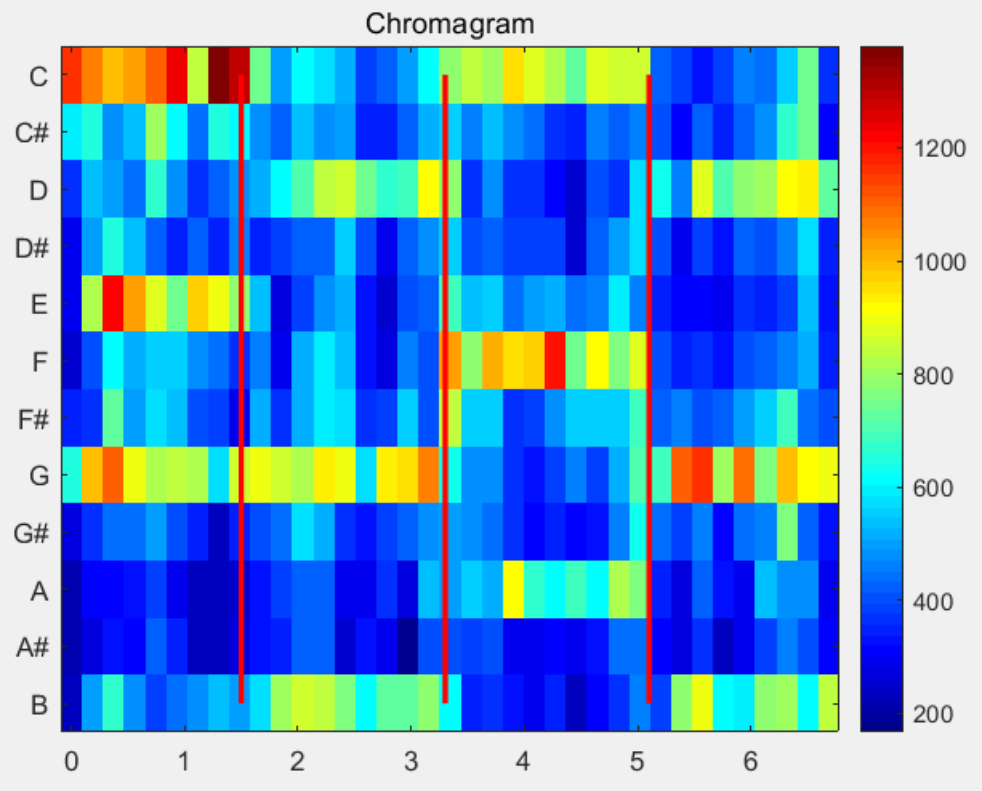
古典吉他音域为大字组 E 到小字二组 b2。注意到吉他演奏多为节奏型伴奏，Adam M. Stark 等研究者已经指出节奏型伴奏主要使用乐器的低音域[3]，所以该算法考虑从大字组 C 开始的$\phi = 1,2$两个 8 度。

实际乐器的高次谐波频率并不总是严格地 k 倍于基频，为了减少谐波失真带来的影响，算法只考虑$h = 1,2$谐波，更高次谐波对算法不会产生影响。

如下为 C，G，F，G 和弦的吉他演奏片段



对该片段计算色谱图，结果如下：



已知 C 和弦组成音为 C、E、G；G 和弦组成音为 G、B、D；F 和弦组成音为 F、A、C。分析色谱图结果，发现和弦组成音的能量明显高于其它音，结果符合预期。

4. 和弦分类器（Classifier）

和弦识别算法是建立在上节所述的色谱图(Chromagram)的基础之上的。因为和弦是由若干音同时弹响组成的，所以色谱图上一定会出现这些音对应的峰值。可以通过某种模型将这些峰值映射到一个确定的和弦，这便是和弦识别的基本思路。

这是分类模型，即将给定的色谱向量归类到所有可能和弦中的一种，具体实现为一个分类器（Classifier），本章详细介绍。

4.1 和弦分类的最近邻(NN)算法

分类器基于最近邻算法（NN）。特征空间由色谱向量构成。首先，所有可能和弦对应的色谱向量都可以预先计算，把预计算的理想色谱向量$\boldsymbol{T_i}$作为样本；其次，把由用户演奏音频计算的色谱向量$\boldsymbol{C}$作为测试样本。于是根据 p-范数计算样本距离为：

$$\delta_i = (\sum_{n=0}^{11} |\boldsymbol{C}(n) - \boldsymbol{T}_i(n)|^p)^{\frac{1}{p}}$$

最近邻算法寻找使δ_i值最小的样本$\boldsymbol{T_i}$，并将该样本对应和弦作为分类结果。

显然，当色谱向量中某一和弦的能量越大时，分类结果越可能得到该和弦，因此这种算法最大化了和弦能量。但在实际运用中，色谱图上的噪声能量也会被当作和弦的一部分考虑，对和弦分类造成干扰。

Adam M. Stark 等研究者从一种逆向的角度给出了改进算法 [3]，该算法并非最大化和弦组成音的能量。相反地，它考虑非和弦组成音的能量（残余能量），并最小化残余能量。

在改进算法下，样本距离计算方法修改为：

$$\delta_i = \frac{\sqrt{\sum_{n=1}^{11}(1 - \boldsymbol{T}_i(n))\boldsymbol{C}(n)^2}}{(12 - N_i)}，\ 1 \le N_i \le 12$$

其中N_i为样本$\boldsymbol{T_i}$所包含的和弦组成音符的个数。系数$1/(12 - N_i)$的目的是避免同等条件下，和弦组成音数目影响样本距离，造成各和弦机会不均等。

最近邻算法寻找样本$\boldsymbol{T_i}$使δ_i值最小化，并将该样本对应的和弦 i 作为分类结果。

4.2 和弦样本计算

根据基础乐理，和弦中每个组成音之间都具有固定的音程关系，所以和弦的理想色谱向量$\boldsymbol{T_i}$是常量，可以预先计算。

本算法将 108 种常用和弦作为样本。下表仅按和弦后缀分类：

和弦后缀	组成音个数	音程关系(半音)
(major)	3	0 4 7
m	3	0 3 7
dim	3	0 3 6
aug	3	0 4 8
sus2	3	0 2 7
sus4	3	0 5 7
maj7	4	0 4 7 11
min7	4	0 3 7 10
dom7	4	0 4 7 10

对于 x 个音组成的和弦，和弦样本$\boldsymbol{T_i}$满足：

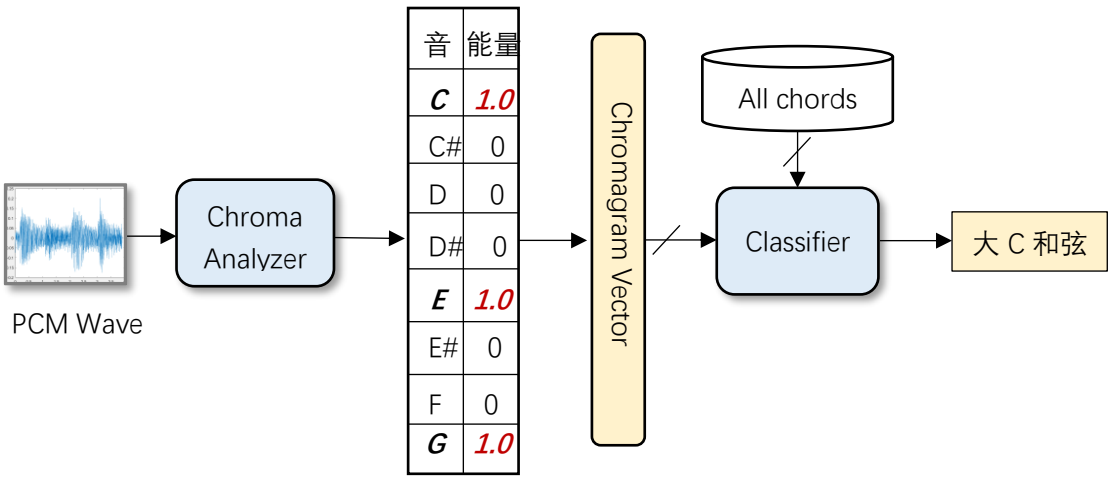
$$\boldsymbol{T_i(n)} = \begin{cases} 1, n = (r + p_i)\ mod\ 12, i = 1\ or\ 2\ or\ ..x \\ 0, 其它 \end{cases}$$

其中p_i为第 i 个组成音相对根音的音程，单位为半音；r 为和弦的根音，$r = 0,1,\ldots,11$。根据上述算法计算的$\boldsymbol{T_i}$即为分类器的样本。

4.3 和弦分类器顶层设计

顶层设计由前端的色谱分析器、后端的和弦分类器构成。输入为用户演奏片段的 PCM 音频信号，通过色谱分析器计算色谱图(Chromagram)，再交由和弦分类器(Classifier)计算用户演奏的和弦，作为算法的输出。

下图为一个实例，显示分类器对一个大 C 和弦分类的流图：



5. 得分计算模型$(\tilde{a}, \tilde{f})$

得分计算模型包括音准得分计算模型$\tilde{a}$、流畅度得分计算模型$\tilde{f}$两部分。

音准得分的关键因素为和弦准确度。基础乐理根据根音、品质、音程三个参数确定一个和弦，因此准确弹奏的和弦必然同时满足上述三个参数的标准值。另一方面，和弦分类器(Classifier)已经计算出用户当前弹奏的和弦，确定了上述三个参数的实际值。评分就是判断实际值与标准值的误差，给出[0,1]区间归一化的分数。

一首音乐作品由 n 个和弦构成，记r_i、q_i、t_i分别表示用户演奏的第 i 个（$1 \le i \le n$）和弦的根音、品质、音程是否与标准演奏 s 相同，满足取 1，不满足取 0，则音准得分$\tilde{a}$可定义为

$$\tilde{a} = \frac{1}{3n}\sum_{i=1}^{n}(r_i + q_i + t_i)\ ,(0 \le \tilde{a} \le 1)$$

流畅度得分的关键评价因素是在规定节拍内是否完成和弦的弹奏。基础乐理中参数 BPM 用于指定音乐绝对速度，含义为每分钟拍数。根据参数 BPM 可计算出每拍的持续时间为：

$$T_b = \frac{BPM}{3600}\ s$$

要确定每个音符的持续时间，还应考虑拍号，如 4/4 拍，3/4 拍等。形式化描述，拍号n/b表示 n 分音符为一拍，而 b 拍构成一个小节。所以，每小节的持续时间计算为：

$$T_s = b \cdot T_b$$

每个全音符的持续时间计算为

$$T_w = n \cdot T_b$$

利用参数T_s、T_w可对标准演奏的和弦时值进行标注。用户演奏的音频经过简单对齐后，与标准演奏保持同一时间线。对于标准演奏中的 n 个和弦，如果在和弦的时值区间内未检测到和弦节奏对应的峰值，说明用户演奏的节奏出现异常。

记h_i表示用户演奏的第 i 个（$1 \le i \le n$）和弦的节奏是否与标准演奏 s 相同，相同取 1，不相同取 0，则流畅度得分$\tilde{f}$可定义为

$$\tilde{f} = \frac{1}{n}\sum_{i=1}^{n} h_i \ , (0 \le \tilde{f} \le 1)$$

综上得到了一组模型$(\tilde{a}, \tilde{f})$，该模型可根据用户演奏的音频参考标准演奏给出 a 值、f 值，进而可通过“算法总述”部分描述的算法计算用户的最终得分。

6. 参考文献

[1] Adam M. Stark. Musicians and Machines: Bridging the Semantic Gap In Live Performance[J]. Queen Mary University of London, 2011

[2] 梅铁民，付天娇，朱向荣. 类谐波积谱基音周期检测算法[J]. 沈阳理工大学学报，2016, 035(002):14-17,23.

[3] Adam M. Stark, Mark D. Plumbley. Real-Time Chord Recognition for Live Performance[J]. ICMC 2009