给定包含N个样本的数据集$D = \{(X_i, Y_i)\}_{i=1}^{N}$,其中$X_i = \{x_t\}_{t=1}^{T}$与$Y_i = \{y_t\}_{t=1}^{T}$为时长为$T$的音频特征序列与指挥动作序列样本，$x_t$和$y_t$分别为第$t$个时间步上的$p$维音频特征$x_t \in R^p$和$q$个关键点的 2 维骨架坐标$y_t \in R^{2q}$。本赛题的任务是在数据集$D$上训练一个映射$G: R^{T\times P} \to R^{T\times 2q}$，生成对应的指挥动作序列$\hat{Y} = G(X)$。

（1） Rhythm Density Error (RDE)

RDE 衡量生成动作与真实动作频率分布的相似度。具体地，首先计算动作的功率谱密度（Power Spectral Density, PSD），再去除极低频的噪音（对应于身体转向、倾斜等低频大幅值动作成分）。本文假设 40BPM 是音乐节奏的一个大致的下界，因此使用$f = 40\text{BPM}/60\text{BPM} \approx 0.7\text{Hz}$作为频率下界，分离小于此界限的动作分量。最后，使用 log 和常数$k = 10^7$将指标值缩放至合适的区间。RDE 的定义如下：

$$\text{RDE}(Y_i, \hat{Y}_i) = \log\left[k\left\|\sum_{j}^{26}\text{PSD}_{f>0.7\text{HZ}}(Y_i[j]) - \sum_{j}^{26}\text{PSD}_{f>0.7\text{HZ}}(\hat{Y}_i[j])\right\|_2^2 + 1\right] \tag{1}$$

（2） Strength Contour Error (SCE)

SCE 用来比对生成动作与真是动作力度变化的相似程度。指挥动作的力度变化可以由各个关键点的一阶差分得到，结果经过一个池化降采样层后，提取到更宽窗口内的力度变化趋势。具体地，对得到的一阶差分之和施加核为 60 帧（2 秒），步长为 30 帧（1 秒）的平均池化，称得到的曲线为力度轮廓（strength contour）。SCE 即是对比生成动作与真实动作力度轮廓之间的差异。类似地，最后使用 log 和常数$k = 10^7$ 将指标值缩放至合适的区间。SCE 的定义如下：

$$\text{SCE}(Y_i, \hat{Y}_i) = \log\left[k\left\|\text{pool}\left(\sum_{j}^{26}Y_i[j]\right) - \text{pool}\left(\sum_{j}^{26}\hat{Y}_i[j]\right)\right\|_2^2 + 1\right] \tag{2}$$

（3） Standard Deviation Percentage (SDP)

SDP 是生成指挥动作与真实指挥动作标准差之比。当生成结果有严重的过度平滑问题时，该动作的标准差将趋于 0%。而理想的生成器生成动作的标准差与真实动作的标准差之比应在 100%附近。令T^y表示动作Y_i的总帧数，$\bar{y}$表示平均姿态，则 SDP 可以定义为：

$$\text{SDP}(Y_i, \hat{Y}_i) = \frac{\text{SD}(\hat{Y}_i)}{\text{SD}(Y_i)}\text{，其中 SD(Y}_\text{i}\text{)} = \sqrt{\sum_{t=1}^{T^y}\frac{\|y_t - \bar{y}\|_2^2}{T^y - 1}} \tag{3}$$