

1. 时域音频信息就是一个点随着时间在振膜垂直方向振动的情况，可表示为一个2D点集，采样率越高，就越接近连续曲线。

sample rate 采样率 = 对这个点所在位置测量的频率，通常就是44100Hz。

bit rate 比特率 = 采样率 \* 量化精度 \* 声道数，是指单位时间内处理的数据量。

buffer size = window size = 每次分析步骤所需的sample数。通常是1024或2048。

hop size = 两个相邻窗口之间错开的sample数，越小，则说明时序解析度越高，计算成本也越高。通常为buffer size的一半或四分之一。

frame size = 帧长，媒体帧的长度。

fps = 帧率。一个帧可能包含多个采样。音频基本都是这样，视频帧则一般一帧一采样。因此fps这个概念通常用于视频和游戏领域。

bit depth = 位深度，每次采样sample里包含的信息的bit数。

channels = 声道数，双声道文件大小是单声道两倍。

2. 简谐波的数学公式如下为 $y = A \sin(\omega t + \varphi)$ ， $A$ 表示振幅（振幅代表信号的音量或响度）， $\omega$ 表示频率， $\varphi$ 表示初始相位。
3. **振幅包络线**(Amplitude Envelope)的目的是提取每一帧的最大振幅并将它们串在一起。具体作法为把信号分解成它的组成窗口，并找出每个窗口内的最大振幅，然后画出每个窗口沿时间的最大振幅。  
 \*\*应用：\*\*可以使用AE进行检测声音是否开始，比如在各种语音处理应用程序中，这可能是某人讲话或外部噪音，而在音乐信息检索（MIR）中，这可能是音符或乐器的开始。  
 \*\*缺点：\*\*AE的主要缺点是对离群值的鲁棒性差。  
 \*\*公式：\*\*计算公式如下，其中 $K$ 表示每帧有多少个样本数。

Max amplitude value of all samples in a frame

$$AE_t = \max_{k=t \cdot K}^{(t+1) \cdot K - 1} s(k)$$

Amplitude envelope at frame  $t$

Last sample of frame  $t$

Amplitude of  $k$ th sample

First sample of frame  $t$

4. **均方根能量**（Root-Mean-Square Energy）：均方根能量表达的是一帧内所有样本点的一个综合信息，与开始检测相反，它尝试感知响度。当我们观察波形时，我们对窗口内的振幅进行平方，然后求和。一

旦完成，我们将除以帧长，取平方根，那将那个窗口的均方根能量。

**应用：**音频分割、音乐流派分类。

**优缺点：**它对于异常值的抵抗力要强得多，这意味着如果我们对音频进行分段，就可以更加可靠地检测到新事件（例如新乐器，某人讲话等）。

**公式：**计算公式如下。

**RMS of all samples in a frame**

$$RMS_t = \sqrt{\frac{1}{K} \cdot \sum_{k=t \cdot K}^{(t+1) \cdot K - 1} s(k)^2}$$

Mean of sum of energy

5. **过零率（ZCR）：**过零速率(ZCR)的目的是研究信号的幅值在每一帧中的变化速率。

**应用：**对于MIR，此功能与识别打击乐器声音有关，因为它们经常具有波动信号，ZCR可以很好地检测到这些声音，并且可以检测到音高。但是，此功能通常用作语音识别中用于语音活动检测的功能。

**公式：**计算公式如下。

$$ZCR_t = \frac{1}{2} \cdot \sum_{k=t \cdot K}^{(t+1) \cdot (K-1)} |\text{sgn}(s(k)) - \text{sgn}(s(k+1))|$$

6. 在