音视频编码原理

本篇博客为面向大众的科普性文章。涉及声音原理、音频文件属性、音频格式等方面。预计阅读时间为10分钟。

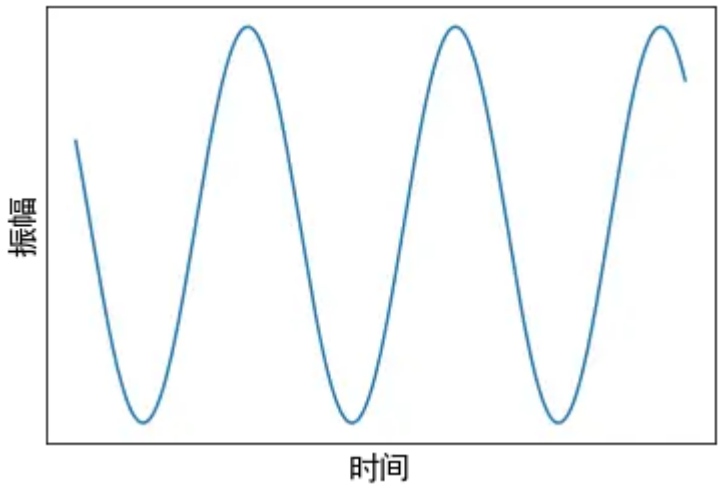
1.何为声音

中学物理中我们知道，声音是物体振动产生的声波。声音通过介质（空气、固体、液体）传入到人耳中，带动听小骨振动，经过一系列的神经信号传递后，被人所感知。

声音是一种波。物体振动时会使介质（如空气）产生疏密变化，从而形成疏密相见的纵波。

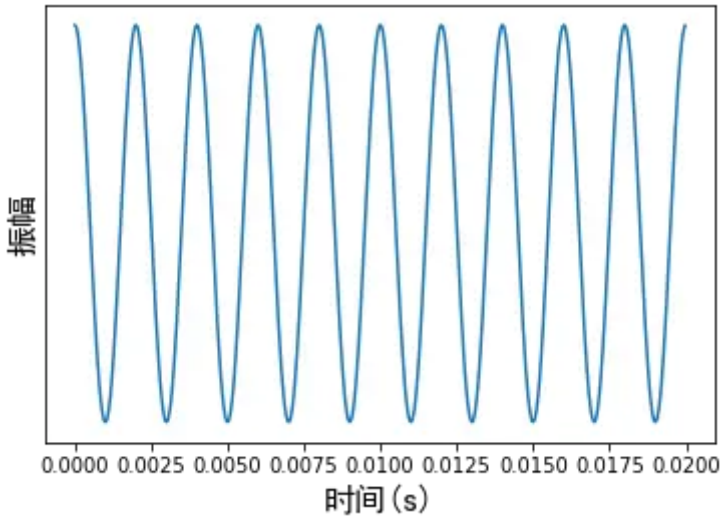
既然声音是波，那么我们就可以用图的形式来表示它。

给定空间中某一点，该点的空气疏密随时间的变化如下：



波形图

下图是一个正弦波，其周期为0.002s，频率为500HZ。



该声音很像视频中的“消音”处理。

频率（音调）：声音1秒内周期性变化的次数

人耳的听觉范围在20Hz-20kHz。 低频的声音沉闷厚重，高频的声音尖锐刺耳。 高于 20kHz的声音为超声波。

振幅（响度）：声音的大小

有的时候，我们用分贝（dB）形容声音大小。值得注意的是，**dB是一个比值，是一个数值，没有任何单位标注。（功率强度之比的对数的10倍）**

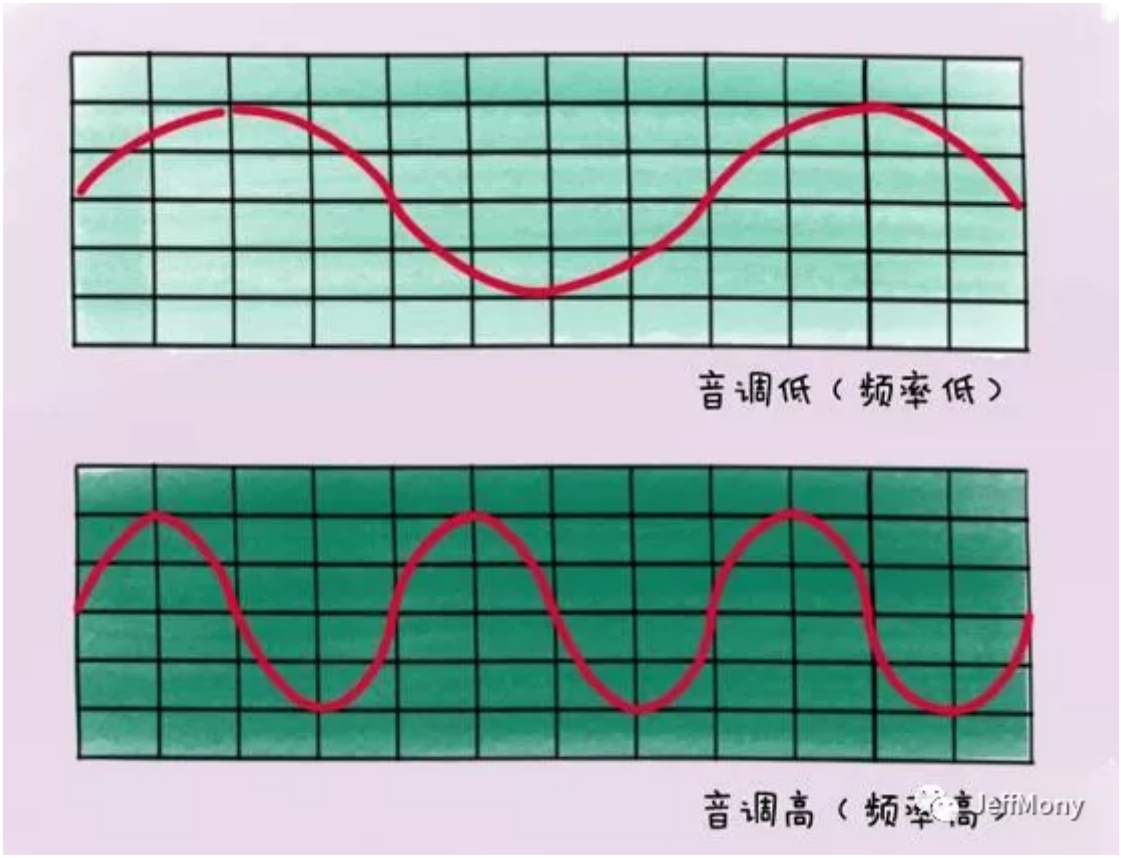
1分贝	刚能听到的声音
15 分贝以下	感觉安静
30 分贝	耳语的音量大小
40 分贝	冰箱的嗡嗡声
60分贝	正常交谈的声音
70分贝	相当于走在闹市区
85分贝	汽车穿梭的马路上
95分贝	摩托车启动声音
100分贝	装修电钻的声音
110分贝	卡拉OK、大声播放MP3 的声音
120分贝	飞机起飞时的声音
150分贝	燃放烟花爆竹的声音

- 音调：声音频率的高低叫做音调(Pitch),是声音的三个主要的主观属性,即音量(响度)、音调、音色(也称音品) 之一。表示人的听觉分辨一个声音的调子高低的程度。音调主要由声音的频率决定,同时也与声音强度有关

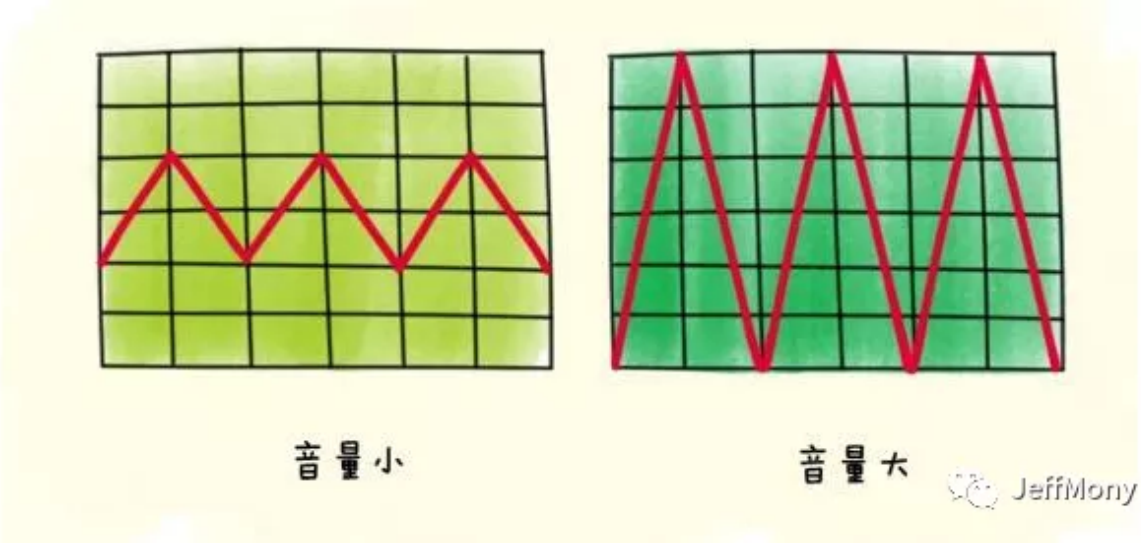
- 音量：人主观上感觉声音的大小（俗称音量），由“振幅”（amplitude）和人离声源的距离决定，振幅越大响度越大，人和声源的距离越小，响度越大。（单位：分贝dB）
-
- 音色：又称声音的品质，波形决定了声音的音色。声音因不同物体材料的特性而具有不同特性，音色本身是一种抽象的东西，但波形是把这个抽象直观的表现。音色不同，波形则不同。典型的音色波形有方波，锯齿波，正弦波，脉冲波等。不同的音色，通过波形，完全可以分辨的。

波长越长

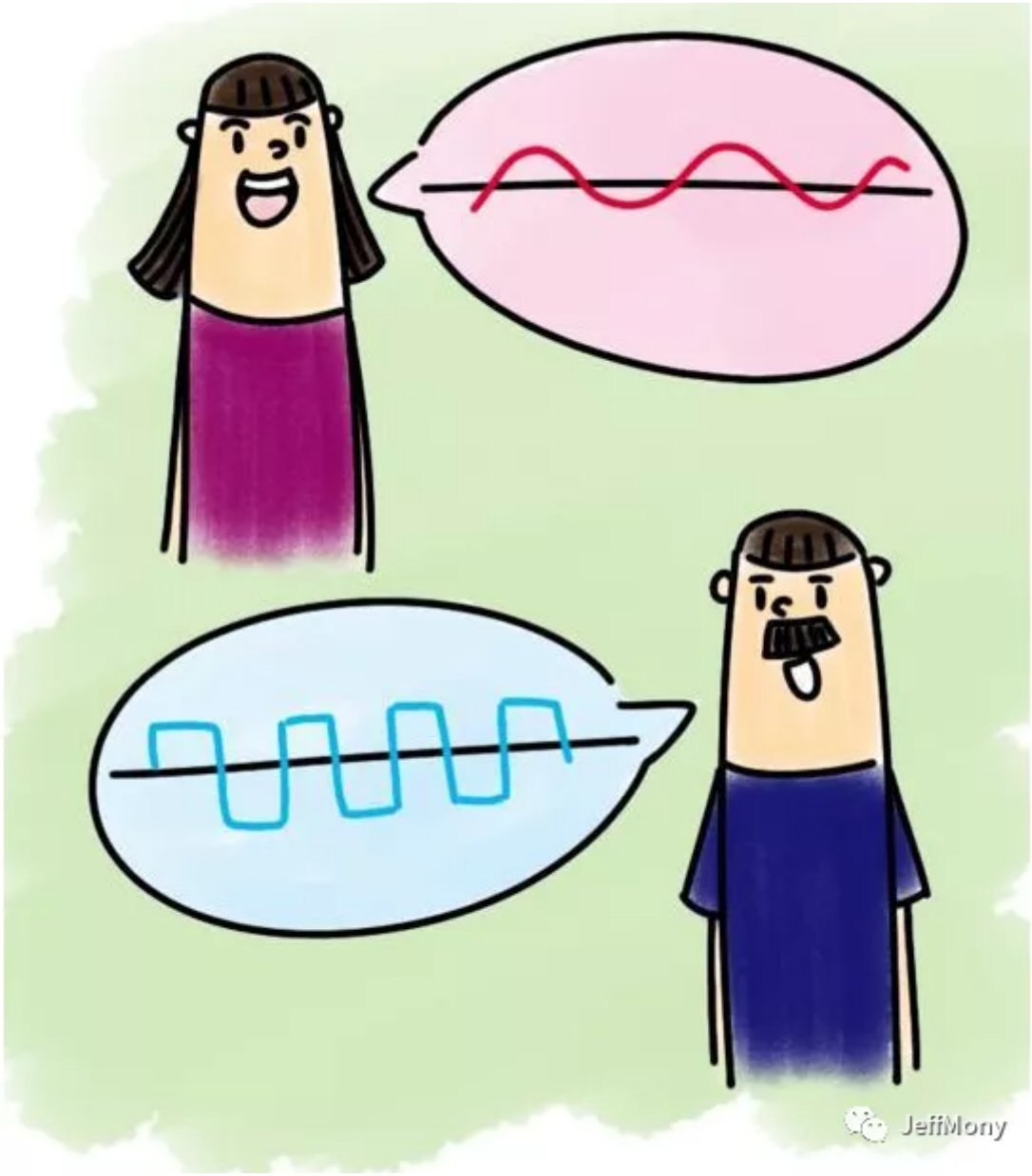
同等条件下，波长（频率）是决定音调高低的因素。



同等条件下，振幅是决定音量高低的因素。



同等条件下，波纹是决定音色因素。



通过上面简单的分析，我们已经知道声音的音量实际上就是由声波的振幅决定的，我们需要调整声波的振幅。播放一个视频，需要经历下面几步：

- 输入视频url
- 确定视频的封装格式
- 开始解封装
- 识别视频的轨道数据
- 分离轨道数据，音频轨道、视频轨道
- **解码视频数据为原始数据，解码音频数据为原始数据**
- 做好音视频同步
- 渲染视频原始数据，播放音频原始数据

通道数 麦克风

上面加黑标红的部分就是我们改变声音振幅的地方，只有将声音数据解码为原始数据，我们加工原始数据的音频流，然后送到AudioTrack或者OpenSL ES内部播放即可。

2.声音采集与存储

采样，指把时间域或空间域的连续量转化成离散量的过程 。

对声音的采样常用麦克风等设备将声音信号转换成电信号，再用模/数转换器将电信号转换成一串用1和0表示的二进制数字（数字信号）。

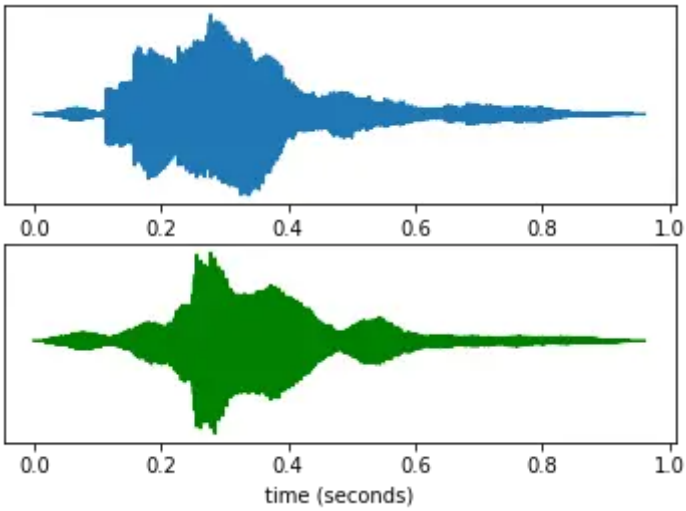
我们每秒对声音采样上万次，获得上万个按照时间顺序排列的二进制数字。于是，我们就将连续变化不断的声音转化成了计算机可储存并识别的二进制数字。

：

该声音由84700个不同的数字组成。 其中的一段数字如下：（二进制数字已转换为十进制）

... 413, 263, 137, 15, -124, -253, -369, -463, -511, -545, -587, -632, -678, -701, -687, -659, -623, -579, -539, -473, -380, -282, -162, -35, 78, 211, 341, 430, 499, 548, 551, ...

如果用图像的形式表示该音频，则图像如下：（横轴是时间，纵轴为振幅，两个图像分别代表左右声道。由于声音频率较大，所以在图像中的信号不是“正弦”，而是实心的。）



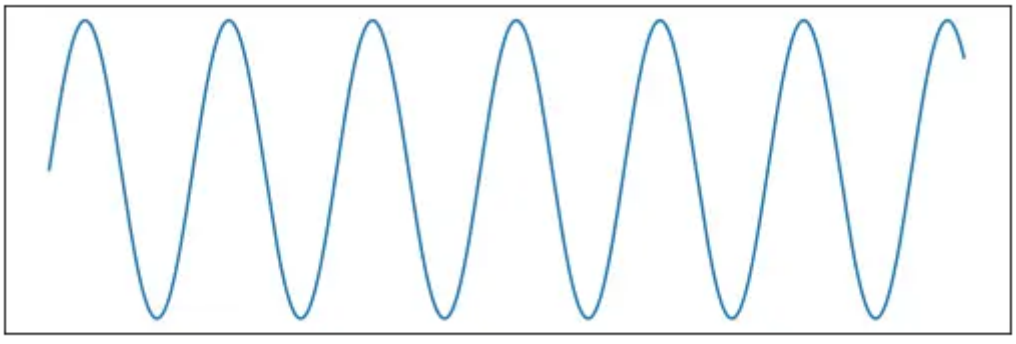
2.1 采样频率

采样频率指录音设备在一秒钟内对声音信号的采样次数。采样频率越高，声音的还原就越真实越自然。

目前主流的采样频率有22.05KHz、44.1KHz、48KHz三种。

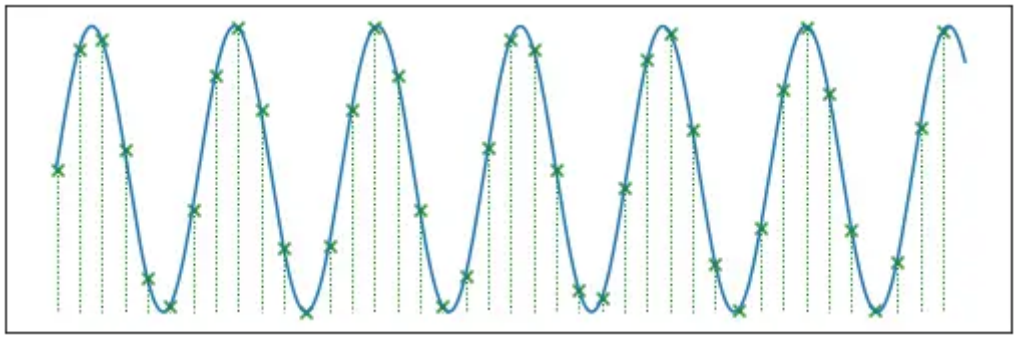
22.05 KHz为FM广播的声音品质，44.1KHz为理论上的CD声音品质。48KHz为人耳可辨别的最高采样频率。

直观理解：一段连续的声音如下

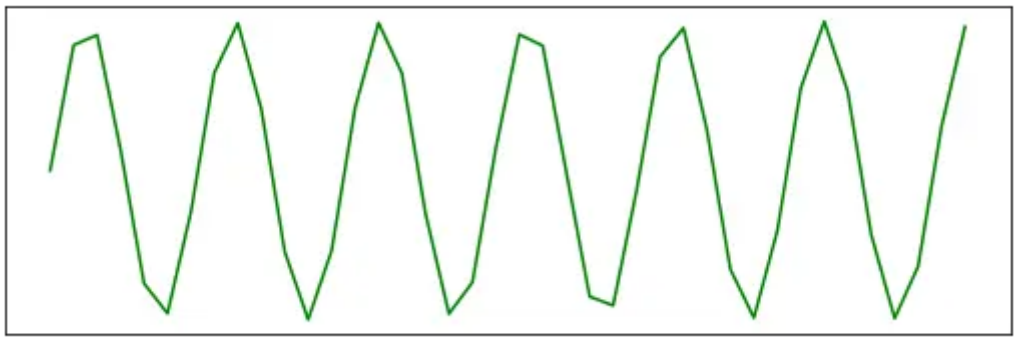


一段声音

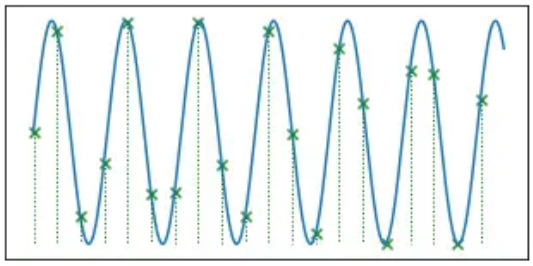
我们等间隔地对其采样

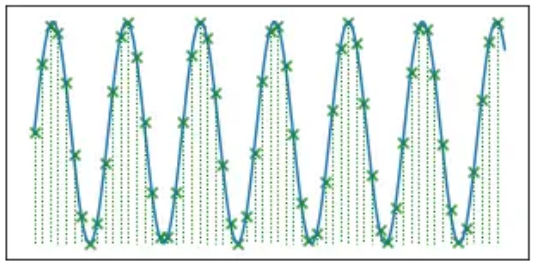


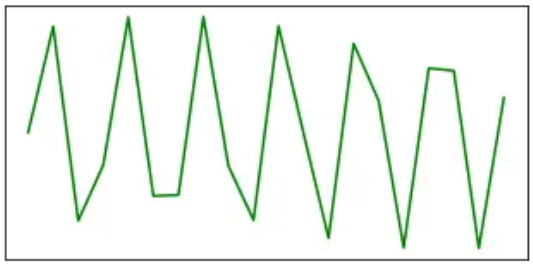
最终，我们真正采样到的音频如下

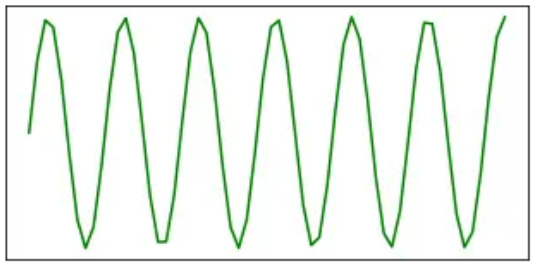


如下图可见，采样频率越高，我们获得的声音品质越好。









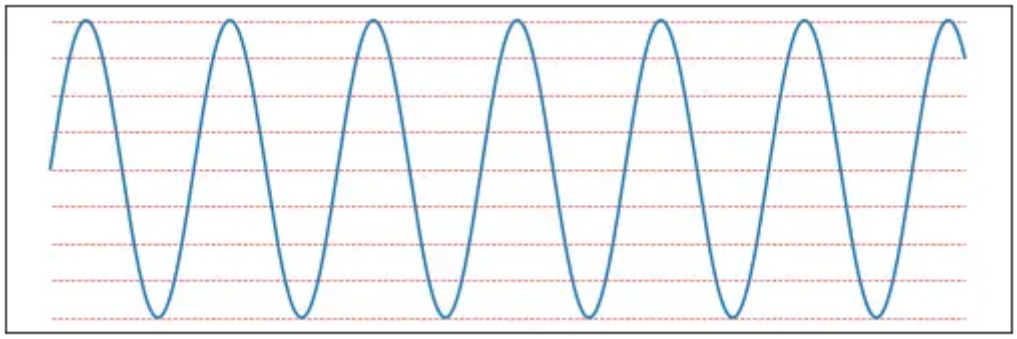
2.2 量化位数

我们不可能获得所有时间下声音的强度，因此声音是等时间间隔、离散采样的。同样，采样获得的数据不可能无限的精确，如数字为63.2222222....，这无法在计算机中储存。因此，采样获得的数据同样也是离散的。

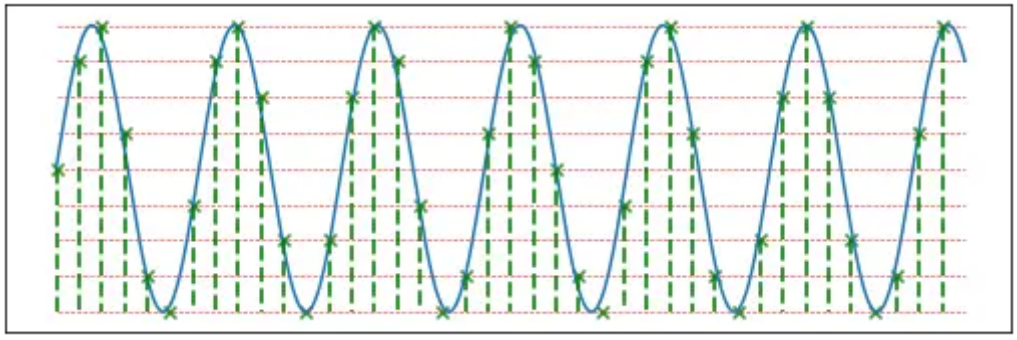
量化位数是音频文件的另一个参数。量化位数越大，声音的质量越高。常用的量化位数有8位、16位和32位。

量化位数指用几位二进制数来存储采样获得的数据。量化位数为8即指用8位二进制数来存储数据，如00010111

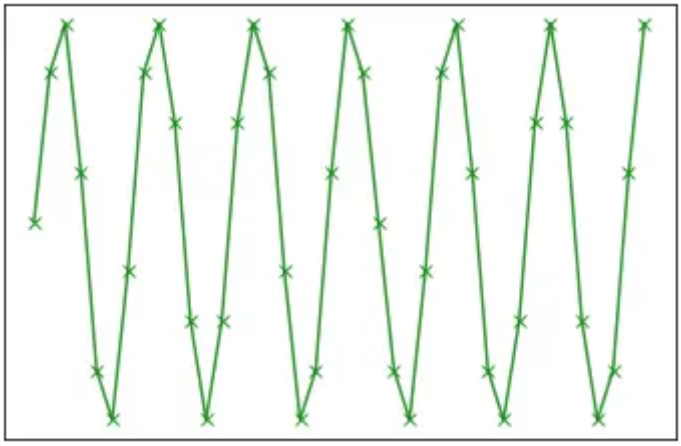
还是之前的例子，有一段正弦声波，假设量化位数为3，即存储的数据只有000/001/010/011/100/101/110/111这8种可能。



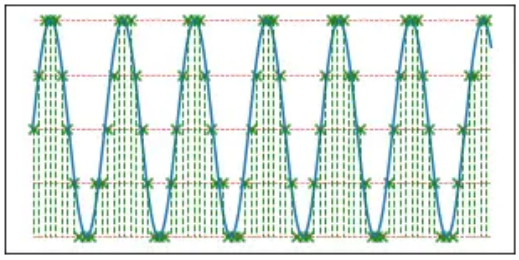
现在，还是等距离采样，不过采样的点只能落在最近的红线上。

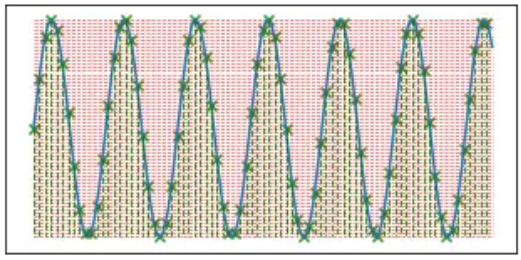


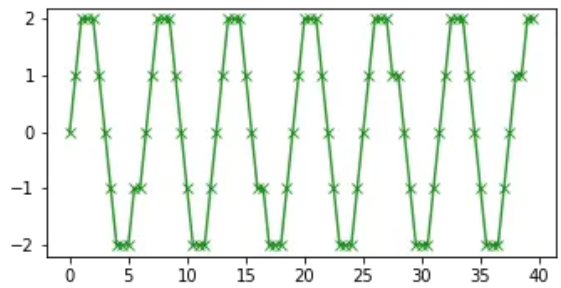
此时，每个点纵坐标的取值只有二的三次方，即只有8中可能。

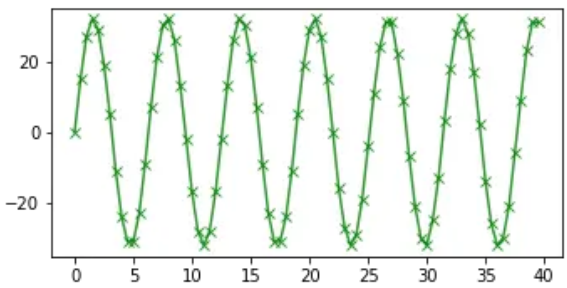


由下图可见，量化位数越大，声音效果越好。









另外值得注意的是，不同量化位数存储的数据不可直接比较。

如4位量化位数存储的1111，其十进制是15，8位量化位数存储的10000000，其十进制是64。不是因为64>15，所以后者对应的声音比前者大。而是应该二者分别除以其总取值范围后在比较。

$$\frac{15}{2^4} > \frac{64}{2^8}$$

前者对应的声音比后者大。

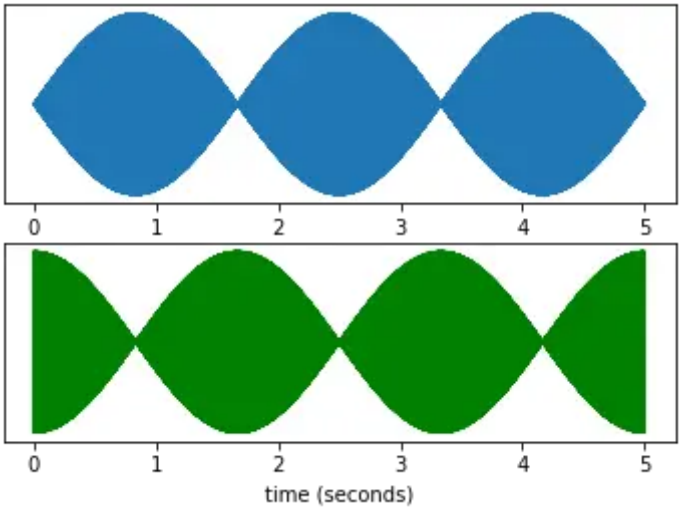
2.3 声道数

声道分为单声道与双声道。

单声道即为左右耳听到的声音相同。

双声道两耳听到的信息不同。相同的声音时间、采样频率和比特率的情况下，双声道文件的存储空间是单声道的两倍。但其会给人空间感，游戏和电影中常采用双声道，可达到“听声辨位”的效果。

示例声音如下：



3 采样频率

> 采样定理在1928年由美国电信工程师H.奈奎斯特首先提出来的，因此称为奈奎斯特采样定理。
>
> 1933年由苏联工程师科捷利尼科夫首次用公式严格地表述这一定理，因此在苏联文献中称为科捷利尼科夫采样定理。
>
> 1948年信息论的创始人C.E.香农对这一定理加以明确地说明并正式作为定理引用，因此在许多文献中又称为香农采样定理。

奈奎斯特采样定理解释了采样率和所测信号频率之间的关系。 阐述了采样率fs必须大于被测信号感兴趣最高频率分量的两倍。

该频率通常被称为奈奎斯特频率fN。即：

奈奎斯特采样定理

```
fs >= 2fH
```

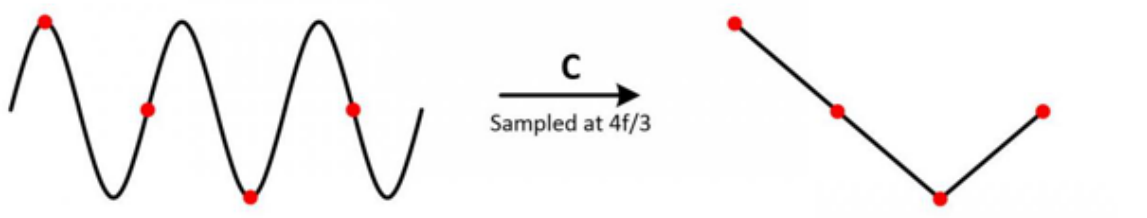
根据奈奎斯特采样定理，需要数字化的模拟信号的带宽必须被限制在采样频率fs的一半以下，否则将会产生混叠效应，信号将不能被完全恢复。这就从理论上要求一个理想的截频为fs/2的低通滤波器。实际中采用的通频带为0~fs/2的低通滤波器不可能既完全滤掉高于的fs/2的分量又不衰减接近于fs/2的有用分量。因此实际的采样结果也必然与理论上的有差别。如果采用高于fs的采样频率，如图1中为2fs，则可以很容易用模拟滤波器先滤掉高于1.5fs的分量，同时完整保留有用分量。采样后混入的界于0.5fs~1.5fs之间的分量可以很容易用数字滤波器来滤掉。这样输入模拟滤波器的设计将比抗混叠滤波器简单的多。

为更好理解其原因，让我们来看看不同速率测量的正弦波。

1. 假设 fS = fN

可以看出，无论我们从哪一点开始采样，每次采集到的数据都是一样的，对应的频率成分为0Hz。

2. 假设 fS = (4/ 3) * fN



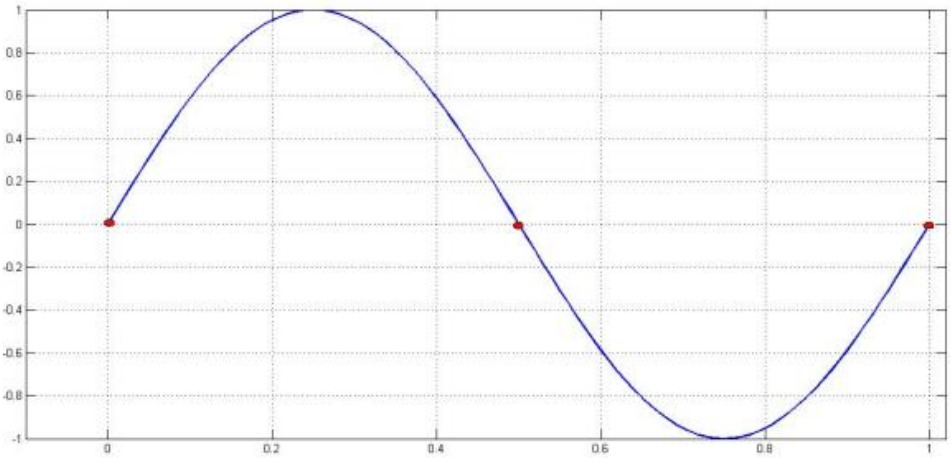
以上采样到的曲线仍然无法还原原有波形的样子。

3. 假设 fS = 2 * fN

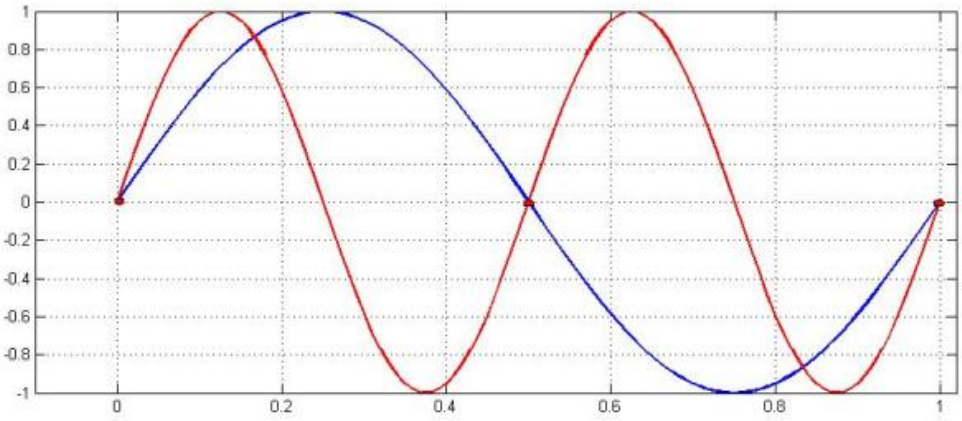
如上图，将这些采样点连成线条，得到的信号形状为三角波，虽然信号的频率成分没有失信，但是很难保证信号的幅值不失真。因为这两个采样点很难位于正弦信号的波峰与波谷处。也就是说，在很大程度上，采样后的信号的幅值是失真的。

我们再考虑如下情况：

假设一条正弦曲线为sin(2π/t)，频率为1Hz。我们以2Hz的频率对该曲线进行采样（每隔0.5s），可以得到3个红色采样数据，如下图：



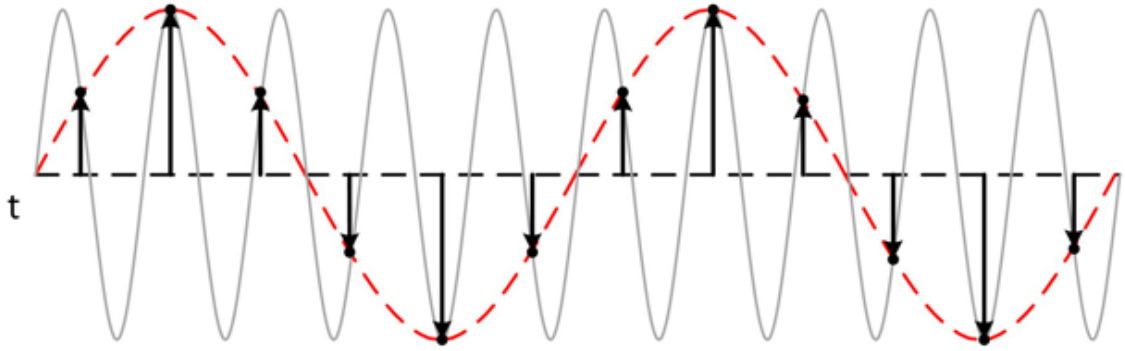
对于这三个点，我们不能确定它对应的正弦曲线是sin(2π/t)，因为sin(4π/t)等倍频曲线也会穿过这三个红色采样点：



混叠

如果信号的采样率低于两倍奈奎斯特频率，采样数据中就会出现虚假的低频成分。 这种现象便称为混叠。

下图显示了800 kHz正弦波1MS/s时的采样。虚线表示该采样率时记录的混叠信号。 800 kHz频率与通带混叠，错误地显示为200 kHz正弦波。



绝大多数信号都是能够进行傅里叶变换的，就意味着，不管一个信号多么复杂，总可以分解为若干个正（余）弦信号的和，对应了信号的频率分量。因此，Nyquist采样定理只需找到信号最大的频率分量，再用2倍于最大频率分量的采样频率对信号进行采样，从理论上解决了，离散信号能够重建出连续信号的问题。