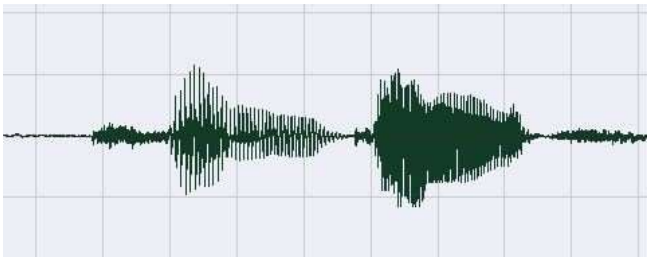
详解语音处理检测技术中的热点——端点检测、降噪和压缩

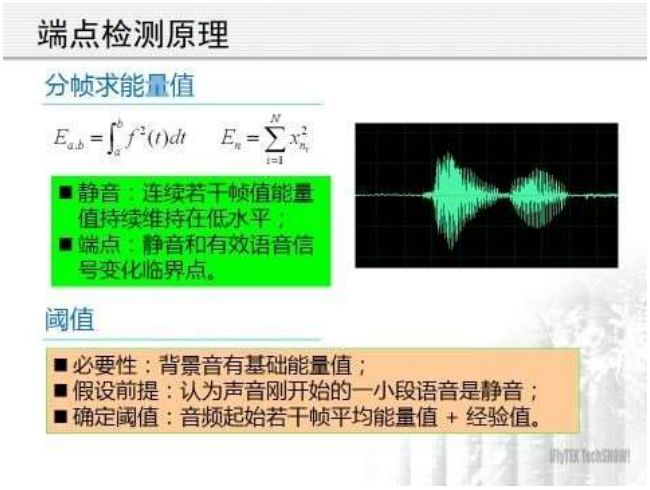
端点检测

首先来看端点检测（Voice Activity Detection, VAD）。音频端点检测就是从连续的语音流中检测出有效的语音段。它包括两个方面，检测出有效语音的起始点即前端点，检测出有效语音的结束点即后端点。

在语音应用中进行语音的端点检测是很必要的，首先很简单的一点，就是在存储或传输语音的场景下，从连续的语音流中分离出有效语音，可以降低存储或传输的数据量。其次是在有些应用场景中，使用端点检测可以简化人机交互，比如在录音的场景中，语音后端点检测可以省略结束录音的操作。



为了能更清楚说明端点检测的原理，先来分析一段音频。上图是一段只有两个字的简单音频，从图上可以很直观的看出，首尾的静音部分声波的振幅很小，而有效语音部分的振幅比较大，一个信号的振幅从直观上表示了信号能量的大小：静音部分能量值较小，有效语音部分的能量值较大。语音信号是一个以时间为自变量的一维连续函数，计算机处理的语音数据是语音信号按时间排序的采样值序列，这些采样值的大小同样表示了语音信号在采样点处的能量。



采样值中有正值和负值，计算能量值时不需要考虑正负号，从这个意义上看，使用采样值的绝对值来表示能量值是自然而然的想法，由于绝对值符号在数学处理上不方便，所以采样点的能量值通常使用采样值的平方，一段包含N个采样点的语音的能量值可以定义为其中各采样值的平方和。

这样，一段语音的能量值既与其中的采样值大小有关，又与其中包含的采样点数量有关。为了考察语音能量值的变化，需要先将语音信号按照固定时长比如20毫秒进行分割，每个分割单元称为帧，每帧中包含数量相同的采样点，然后计算每帧语音的能量值。

如果音频前面部分连续M0帧的能量值低于一个事先指定的能量值阈值E0，接下来的连续M0帧能量值大于E0，则在语音能量值增大的地方就是语音的前端点。同样的，如果连续的若干帧语音能量值较大，随后的帧能量值变小，并且持续一定的时长，可以认为在能量值减小的地方即是语音的后端点。

现在的问题是，能量值阈值E0怎么取？M0又是多少？理想的静音能量值为0，故上面算法中的E0理想状态下取0。不幸的是，采集音频的场景中往往有一定强度的背景音，这种单纯的背景音当然算静音，但其能量值显然不为0，因此，实际采集到的音频其背景音通常有一定的基础能量值。

我们总是假设采集到的音频在起始处有一小段静音，长度一般为几百毫秒，这一小段静音是我们估计阈值E0的基础。对，总是假设音频起始处的一小段语音是静音，这一点假设非常重要！！！！在随后的降噪介绍中也要用到这一假设。在估计E0时，选取一定数量的帧比如前100帧语音数据（这些是“静音”），计

算其平均能量值，然后加上一个经验值或乘以一个大于1的系数，由此得到E0。这个E0就是我们判断一帧语音是否是静音的基准，大于这个值就是有效语音，小于这个值就是静音。

至于M0，比较容易理解，其大小决定了端点检测的灵敏度，M0越小，端点检测的灵敏度越高，反之越低。语音应用的场景不同，端点检测的灵敏度也应该被设置为不同的值。例如，在声控遥控器的应用中，由于语音指令一般都是简单的控制指令，中间出现逗号或句号等较长停顿的可能性很小，所以提高端点检测的灵敏度是合理的，M0设置为较小值，对应的音频时长一般为200-400毫秒左右。在大段的语音听写应用中，由于中间会出现逗号或句号等较长时间的停顿，宜将端点检测的灵敏度降低，此时M0值设置为较大值，对应的音频时长一般为1500-3000毫秒。所以M0的值，也就是端点检测的灵敏度，在实际中应该做成可调整的，它的取值要根据语音应用的场景来选择。

以上只是语音端点检测的很简单的一般原理，实际应用中的算法远比上面讲的要复杂。作为一个应用较广的语音处理技术，音频端点检测仍然是一个较为活跃的研究方向。科大讯飞已经使用循环神经网络（Recurrent Neural Networks, RNN）技术来进行语音的端点检测，实际的效果可以关注讯飞的产品。

降噪

降噪又称噪声抑制（Noise Reduction），前文提到，实际采集到的音频通常会有一定强度的背景音，这些背景音一般是背景噪音，当背景噪音强度较大时，会对语音应用的效果产生明显的影响，比如语音识别率降低，端点检测灵敏度下降等，因此，在语音的前端处理中，进行噪声抑制是很有必要的。

噪声有很多种，既有频谱稳定的白噪声，又有不稳定的脉冲噪声和起伏噪声，在语音应用中，稳定的背景噪音最为常见，技术也最成熟，效果也最好。本课程只讨论稳定的白噪声，即总是假设背景噪声的频谱是稳定或者是准稳定的。

前面讲的语音端点检测是在时域上进行的，降噪的过程则是在频域上进行的，为此，我们先来简单介绍或者说复习一下用于时域-频域相互转换的重要工具——傅里叶变换。

为了更容易理解，先看高等数学中学过的傅里叶级数，高等数学理论指出，一个满足Dirichlet条件的周期为2T的函数f(t)，可以展开成傅里叶级数：

$$f(t)=\frac{a_0}{2}+\sum_{n=1}^{+\infty}[a_n\cos(n\omega t)+b_n\sin(n\omega t)]$$

其中 $\omega=\frac{\pi}{T}$，三角系数 a_n 和 b_n 分别为：

$$\left.\begin{aligned}a_n&=\frac{1}{T}\int_{-T}^{T}f(t)\cos(n\omega t)dt, n=0,2,4,\cdots\\ b_n&=\frac{1}{T}\int_{-T}^{T}f(t)\sin(n\omega t)dt, n=1,3,5,\cdots\end{aligned}\right\}$$

对于一般的连续时域信号f(t)，设其定义域为[0,T]，对其进行奇延拓后，其傅里叶级数如下式：

$$f(t)=\sum_{n=1}^{\infty}[b_n\sin(n\omega t)],\quad t\in[0,T]$$

bn的计算同上，由上式可知，任何一个连续的时域信号f(t)，都可以由一组三角函数线性叠加而成。或者说， f(t)都可以由一个三角函数线性组合组成的序列来无限的逼近。信号的傅里叶级数展示的是构成信号的频率以及各个频率处的振幅，因此，式子的右端又可以看做是信号f(t)的频谱，说的更直白一点，信号

的频谱就是指这个信号有哪些频率成分，各个频率的振幅如何。上式从左到右的过程是一个求已知信号的频谱的过程，从右到左的过程是一个由信号的频谱重构该信号的过程。

虽然由信号的傅里叶级数很容易理解频谱的概念，但在实际中求取信号的频谱时，使用的是傅里叶级数的一种推广形式——傅里叶变换。

傅里叶变换是一个大的家族，在不同的应用领域，有不同的形式，在这里我们只给出两种形式——连续形式的傅里叶变换和离散傅里叶变换：

一般形式的 Fourier 变换　　序列 $\{x_n\}_{n=0}^{N-1}$ 的离散 Fourier 变换：

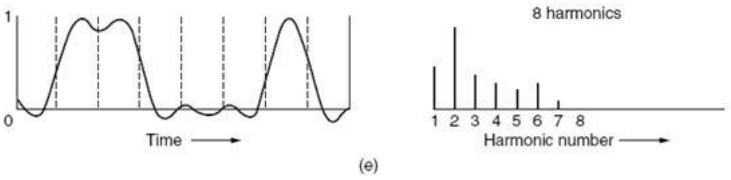
$$F(\omega)=\int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt \qquad X(k)=\sum_{n=0}^{N-1}x_n e^{-j\frac{2\pi kn}{N}}$$

其中的j是虚数单位，也就是j*j=-1，其对应的傅里叶逆变换分别为：

$$F(\omega)=\int_{-\infty}^{+\infty} f(t)e^{-j\omega t}dt \qquad x_n=\frac{1}{N}\sum_{k=0}^{N-1}X[k]e^{\frac{j2\pi kn}{N}}$$

在实际应用中，将数字采样信号进行傅里叶变换后，可以得到信号的频谱。频域上的处理完成后，可以使用傅里叶逆变换将信号由频域转换到时域中。对，傅里叶变换是一个可以完成由时域向频域转换的重要工具，一个信号经傅里叶变换后，可以得到信号的频谱。

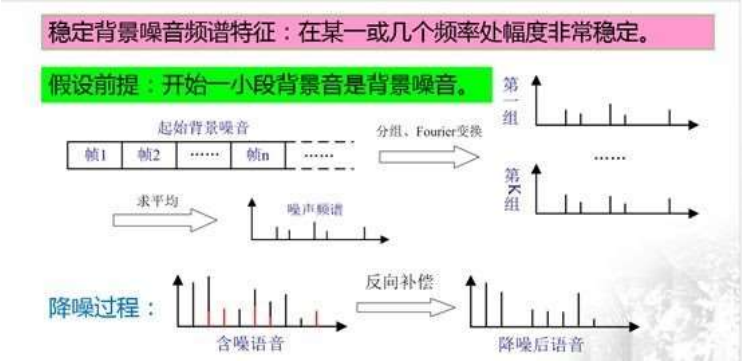
以上是傅里叶变换的简单介绍，数学功底不太好的朋友看不大懂也没关系，只要明白，一个时域信号进行傅里叶变换后，可以得到这个信号的频谱，即完成如下转换：



左面的是时域信号，右面的是对应的频谱，时域信号一般关注的是什么时间取什么值，频域信号关心的是频率分布和振幅。

有了以上的理论作为基础，理解降噪的原理就容易多了，噪音抑制的关键是提取出噪声的频谱，然后将含噪语音根据噪声的频谱做一个反向的补偿运算，从而得到降噪后的语音。这句话很重要，后面的内容都是围绕这句话展开的。

噪声抑制的一般流程如下图所示：

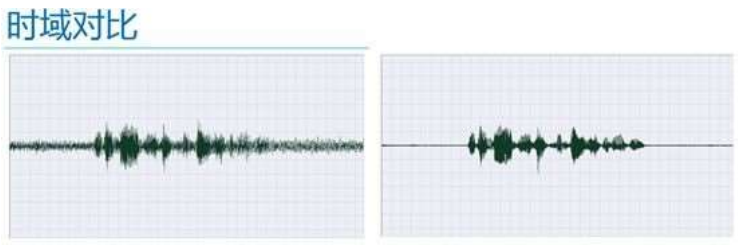


同端点检测类似，假设音频起始处的一小段语音是背景音，这一假设非常重要，因为这一小段背景音也是背景噪声，是提取噪声频谱的基础。

> 降噪过程：首先将这一小段背景音进行分帧，并按照帧的先后顺序进行分组，每组的帧数可以为10或其他值，组的数量一般不少于5，随后对每组背景噪声数据帧使用傅里叶变换得到其频谱，再将各频谱求平均后得到背景噪声的频谱。

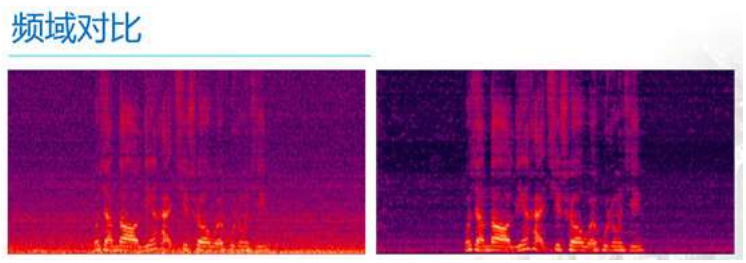
得到噪声的频谱后，降噪的过程就非常简单了，上图下面左侧的图中红色部分即为噪声的频谱，黑色的线为有效语音信号的频谱，两者共同构成含噪语音的频谱，用含噪语音的频谱减去噪音频谱后得到降噪后语音的频谱，再使用傅里叶逆变换转回到时域中，从而得到降噪后的语音数据。

下图展示了降噪的效果



左右两幅图是降噪前后时域中的对比，左面的是含噪语音信号，从图中可以看到噪声还是很明显的。右侧的是降噪后的语音信号，可以看出，背景噪声被大大的抑制了。

下面两幅图是频域中的对比



其中横轴表示时间轴，纵轴表示频率，左面的是含噪语音，其中的亮红色部分是有效语音，而那些像沙子一样的紫色的部分则是噪声。从图中可以看出，噪声不仅是“无时不在”，而且还是“无处不在”，也就是在各种频率处都有分布，右侧的是降噪后的语音，可以很明显的看出，降噪前那些像沙子一样的紫色的部分淡了很多，就是噪声被有效的抑制了。

在实际应用中，降噪使用的噪声频谱通常不是一成不变的，而是随着降噪过程的进行被持续修正的，即降噪的过程是自适应的。这样做的原因一方面是语音数据前部的静音长度有时不够长，背景噪声数据不足导致得到的噪声频谱往往不够准确，另一方面，背景噪声往往不是绝对稳定的，而是渐变的甚至会突变到另一种稳定的背景噪声。

这些原因都要求在降噪的过程中对使用的噪声频谱做及时修正，以得到较好的降噪效果。修正噪声频谱的方法是使用后继音频中的静音，重复噪声频谱提取算法，得到新的噪声频谱，并将之用于修正降噪所用的噪声频谱，所以降噪的过程中仍然要使用端点检测中用到的如何判断静音。噪声频谱修正的方法或者是新旧频谱进行加权平均，或者使用新的噪声频谱完全替换使用中的噪声频谱。

以上介绍的是降噪的非常简单的原理。实际应用中的降噪算法远比上面介绍的要复杂，现实中的噪声源多种多样，其产生的机理和特性也较为复杂，所以噪声抑制在现今仍然是一个较为活跃的研究领域，各种新技术也层出不穷，比如在实际应用中已经使用了多麦克风阵列来进行噪声抑制。