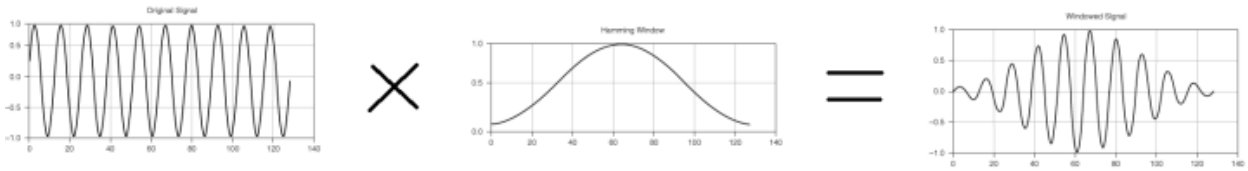
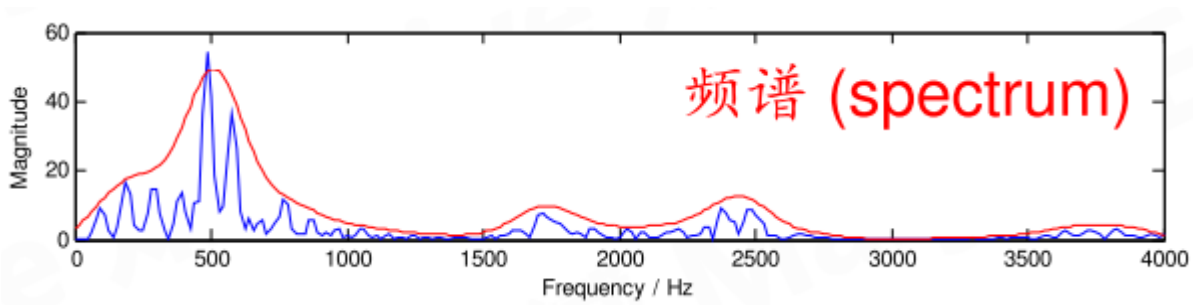
1. 语音中最重要的特征是**各种频率成分的分布**，为此我们使用的数学工具是**傅里叶变换**，傅里叶变换的要求是输入信号是（平稳的），所以对于语音信号，我们从微观上看，在比较短的时间里面，语音信号看作平稳的，就可以截出来做傅里叶变换了。所以需要（**分帧处理**）。
2. 帧的长度需要满足两个条件：
   - 宏观上，要小于一个（**音素**）的长度，即最小的有意义的语音，音素的持续时间大约是（**50毫秒~200毫秒**），所以帧长一般取小于（**50**）毫秒
   - 微观上，它又必须包括足够多的振动周期，因为傅里叶变换是要分析频率的，只要重复足够多次才能分析频率。语音的基频，男声在100赫兹左右，女声在200赫兹左右，换算成周期就是（**10毫秒和5毫秒**）。既然一帧要包含多个周期，所以一般取至少**20毫秒**
   - 所以。帧长一般取为20~50毫秒，20、25、30、40、50都是常用的，
3. 取出来的一帧信号，在做傅里叶变换之前，要先进行【**加窗**】的操作，即与一个窗函数相乘，如下图



   加窗的目的是让一帧信号的幅度在两端渐变到0。渐变对傅里叶变换有好处，可以提高**变换结果**（即频谱）的分辨率。
4. 加窗的代价是一帧信号两端的部分被削弱了，没有像中央的部分那样得到重视。弥补的办法是，帧不要背靠背地截取，而是相互重叠一部分。相邻两帧的起始位置的时间差叫做（**帧移**），常见的取法是取为（**帧长的一半**），或者固定取为（**10毫秒**）
5. 一帧信号做傅里叶变换，得到的结果叫（**频谱**），它就是下图中的蓝线

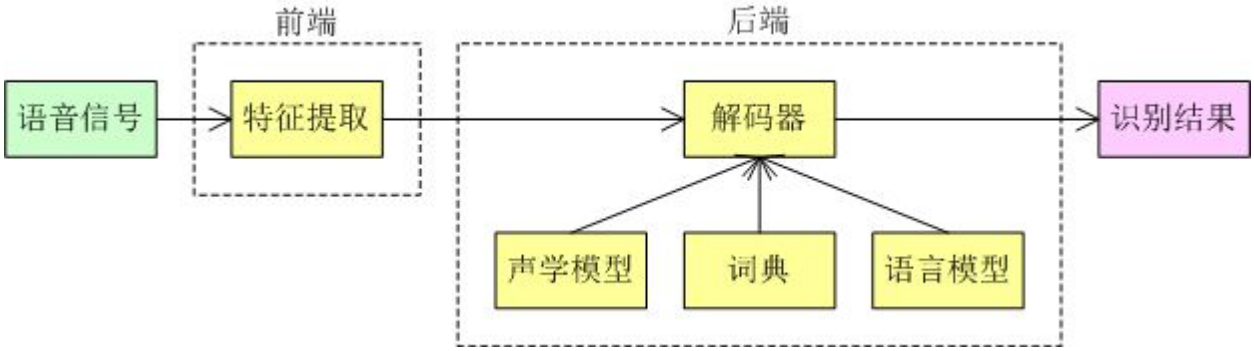


   - 图中的横轴是（**频率**），纵轴为（**幅度**）。
   - 频谱上可以看出这帧语音在480和580Hz附近的能量比较强
   - 语音的频谱，常常呈现出【**精细结构**】和【**包络**】两种模式
     - **精细结构**：蓝线上的小峰，它们在横轴上的间距就是【**基频**】，它体现了语音的【**音高**】-峰越稀疏，基频越高，音高也越高。
     - **包络**：连接这些小峰峰顶的平滑曲线（红线），它代表了口型，即发的是哪个音。包络上的峰叫【**共振峰**】，图中可以看出有四个，分别在500、1700、2450、3800Hz附近。有经验的人，根据共振峰的位置，就能看出发的是什么音。

6. **(采样率）**由人耳能听到的频率范围决定的。人耳能听到的最高频率大约是**（20kHz）**，根据定理，采样率至少要是最高频率的（**2倍**），才能保证不失真。规定，标准的采样率为**（44100Hz）**。实际中为了节省存储空间，也常常用更低的采样率，比如**（22050Hz）（11025Hz）（16000Hz）（8000Hz）**。8kHz是最低的了，再低就不行了
7. **对于wav格式**，比特率是由**（采样率）**、**（采样深度）**、**（声道数）**决定的。
   - [ ] 采样深度：每个样本用几个比特存储，常见的有8或16
   - [ ] 声道数：1或2
   - [ ] 对于4.41kHz的采样率、16比特采样深度、双声道，那么比特率就是

$$44100 * 16 * 2 = 14411.2kbps$$

8. 对于其他格式，比如：mp3，会对数据进行压缩，所以可以用更小的比特率达到同样的音质。
9. 语音识别系统模块，最近几年兴起的神经网络，颠覆了上面框图中的一些模块



10. 首尾端的静音切除，这个操作称为VAD
11. 归一化：

    标准化：matlab中的mapstd，https://blog.csdn.net/hqh45/article/details/42965481
12.