Lecture 02: 语音信号处理

Kai Yu and Yanmin Qian

Cross-media Language Intelligence Lab (X-LANCE)
Department of Computer Science & Engineering
Shanghai Jiao Tong University

2021

目录

- 基本波形处理
- 数字语音波形
- 谱分析
- 听觉系统与听觉特性

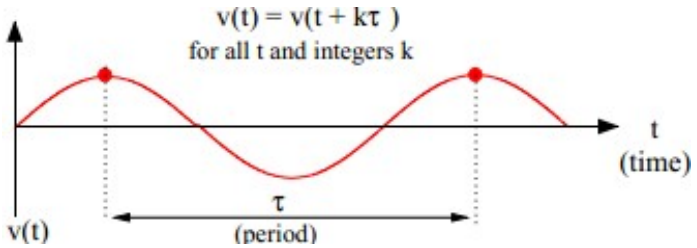
目录

信号的类型

周期性

确定信号: 信号根据已有的公式而产生

周期 信号: 根据周期 τ 进行重复



非周期 信号: 任何没有固定周期的信号

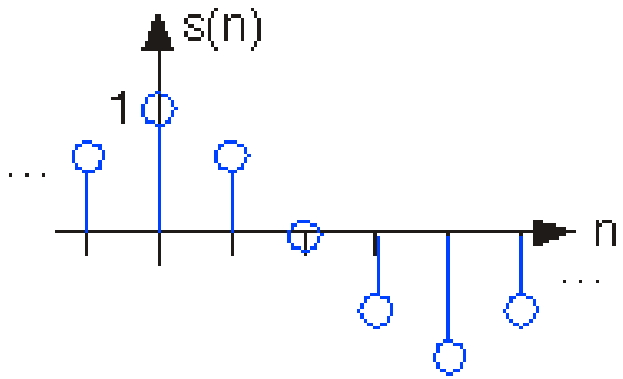
随机信号: 在 t 时刻的信号是一个随机变量的函数: 不可预见的

图:

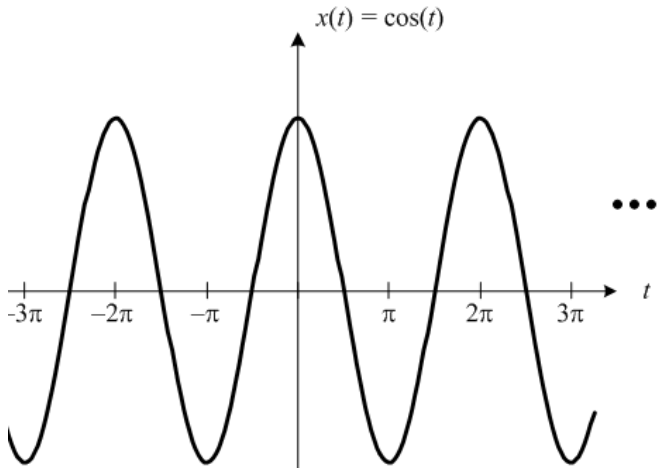
信号的类型

离散和连续: 4 种可能性

离散-时间/幅值信号:



连续-时间/幅值信号:



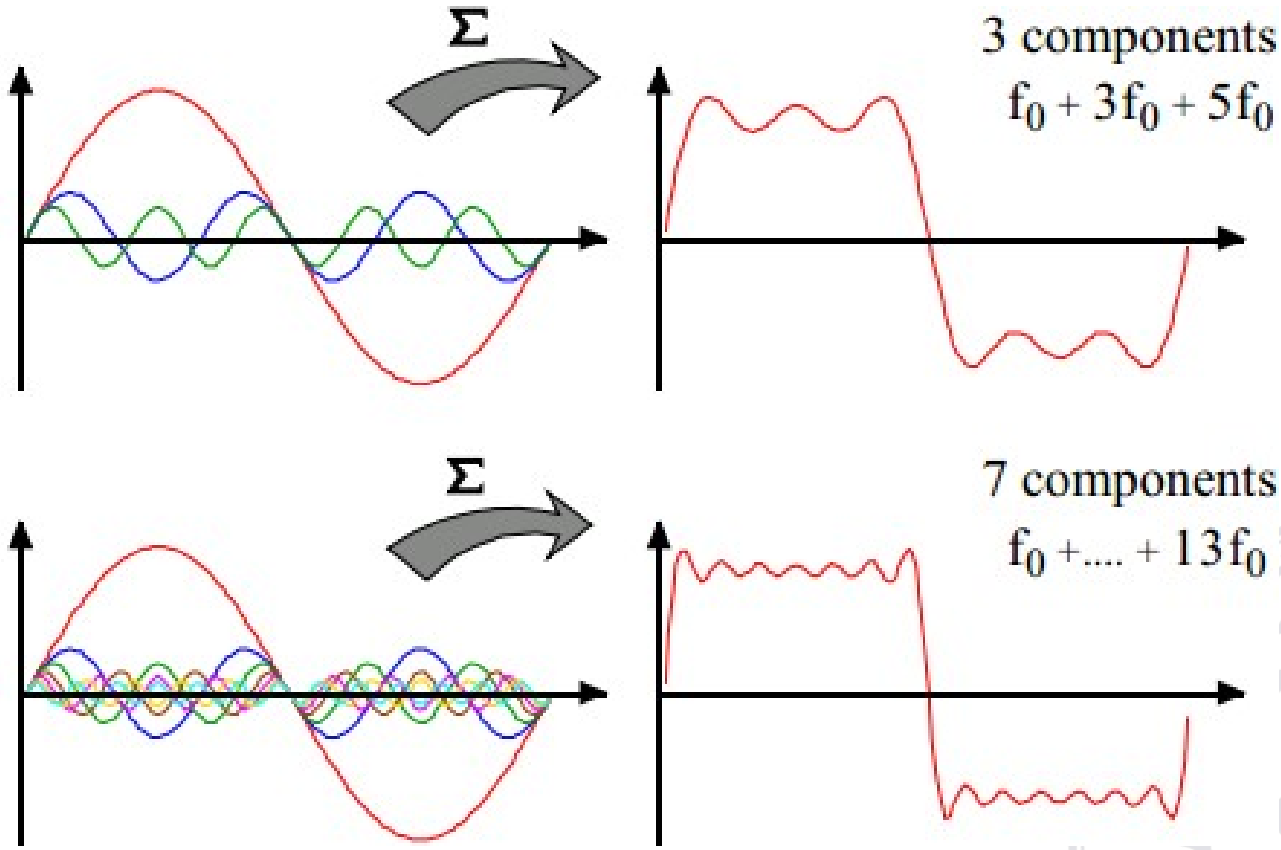
任何频率为 f_0 的周期信号都可以通过具有合适幅值与相位的频率信号 $f_0, 2f_0, 3f_0, 4f_0, 5f_0, ...$ 的叠加来进行精确重构。f_0 被称作为**基本频率 fundamental frequency**，$2f_0, 3f_0$ 等被称为**谐波 harmonics**.

$$s_t = s(t) = \sum_{p=0}^{N-1} A_p \cos(\omega p t + \phi_p)$$

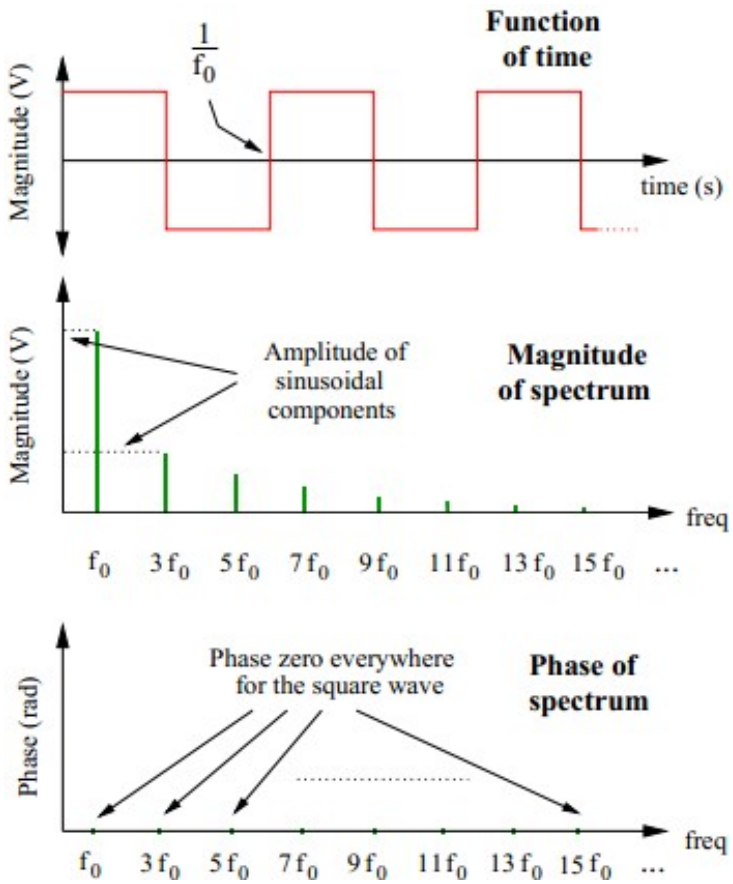
其中 A_p 和 ϕ_p 分别定义为第 p^{th} 次谐波的幅值 和相位.

傅里叶分析

示例: 方波 (仅仅是齐次谐波)



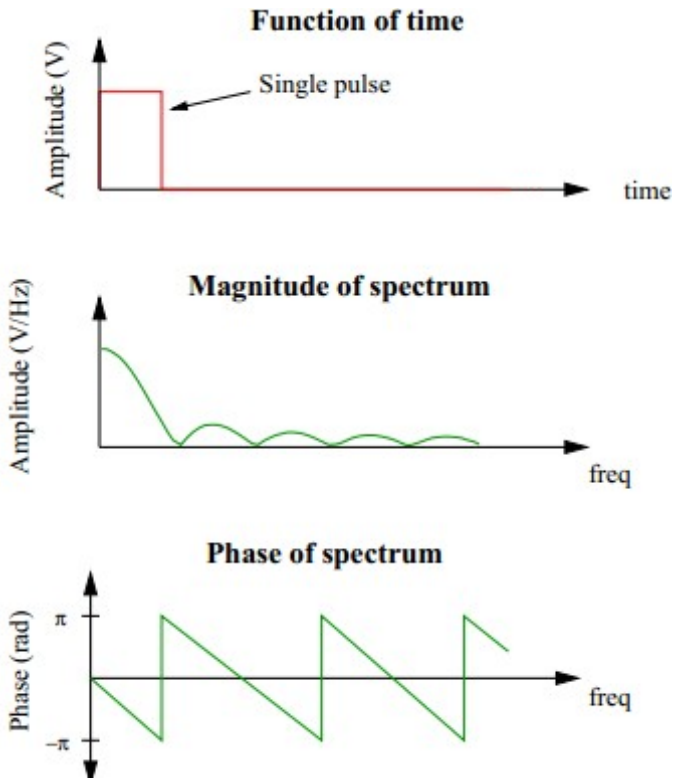
任何周期函数的特性可以用它的谐波分量的幅值和相位来表示。这个特性称为谱特性。



傅里叶分析

非周期信号

- 周期信号仅仅在基频的整数倍的频点上有谱分量
- 非周期信号和随机信号的谱是一个频率的连续函数，i.e. 在所有的频点上都有值



假定 $f(t)$ 是一个连续时间信号，如果有 $\int_{-\infty}^{+\infty} |f(t)|\, dt < \infty$，此信号的**傅里叶变换** 存在，定义为

$$F(jw) = \int_{-\infty}^{+\infty} f(t) e^{-jwt} dt$$

其中 $e^{-jwt} = \cos(wt) + j \sin(wt)$.
此外，**傅里叶反变换** 定义为:

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} F(jw) e^{jwt} dw$$

注意: **周期信号** 可以被展开成傅里叶级数，它的傅里叶变换也就是一些冲击脉冲序列.

离散时间傅里叶变换 DTFT

$$F(e^{jw}) = \sum_{n=-\infty}^{+\infty} f[n]e^{-jwn}$$

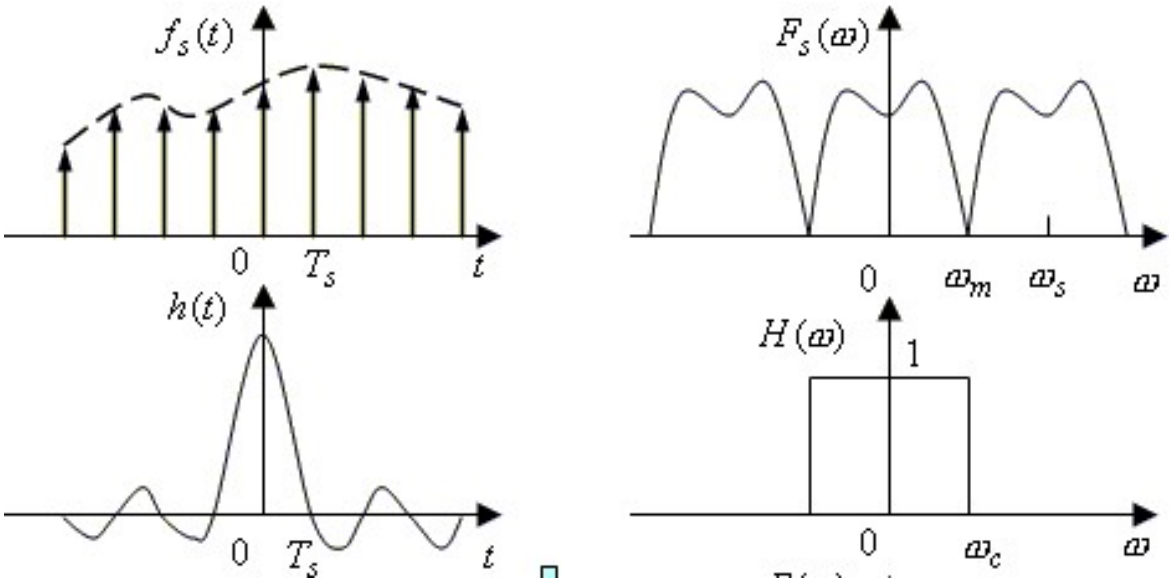
离散时间傅里叶反变换 IDTFT

$$f[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} F(e^{jw})e^{jwn}dw$$

注意: $F(e^{jw})$ 是周期的

$$F(e^{j(w+2\pi)}) = F(e^{jw})$$

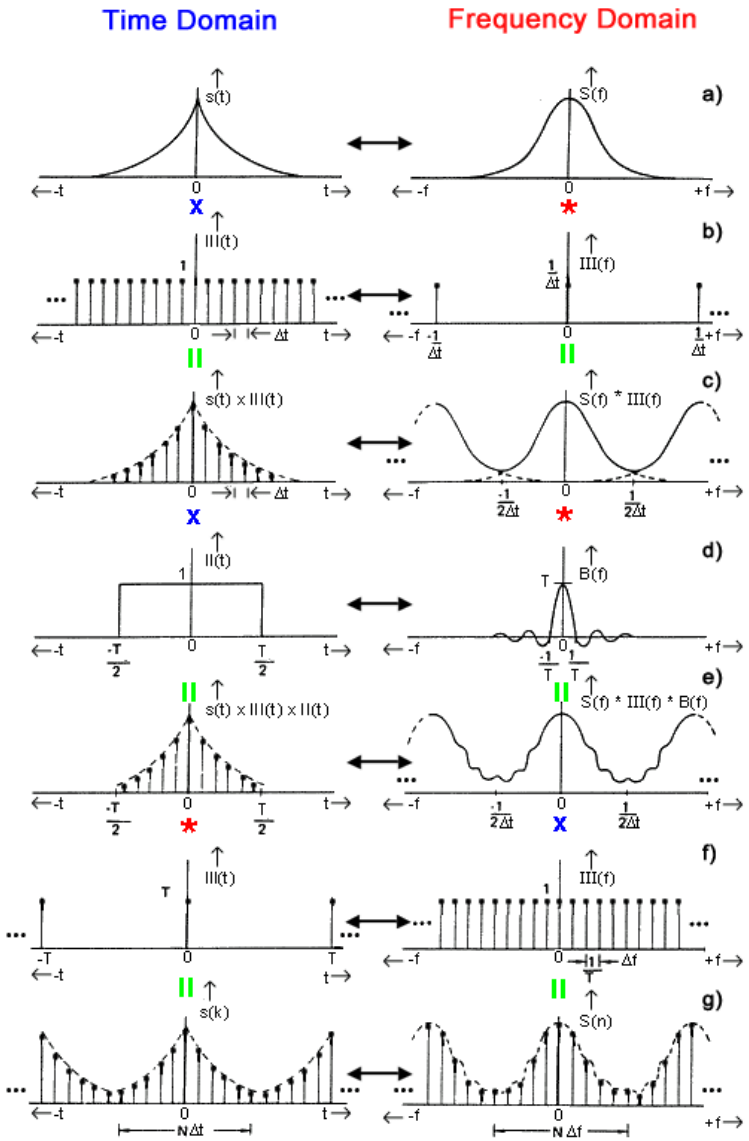
时间域和频域之间的关系



时间	频率
连续 + 非周期	连续 + 非周期
连续 + 周期	离散 + 非周期
离散 + 非周期	连续 + 周期
离散 + 周期	离散 + 周期

离散傅里叶变换 DFT

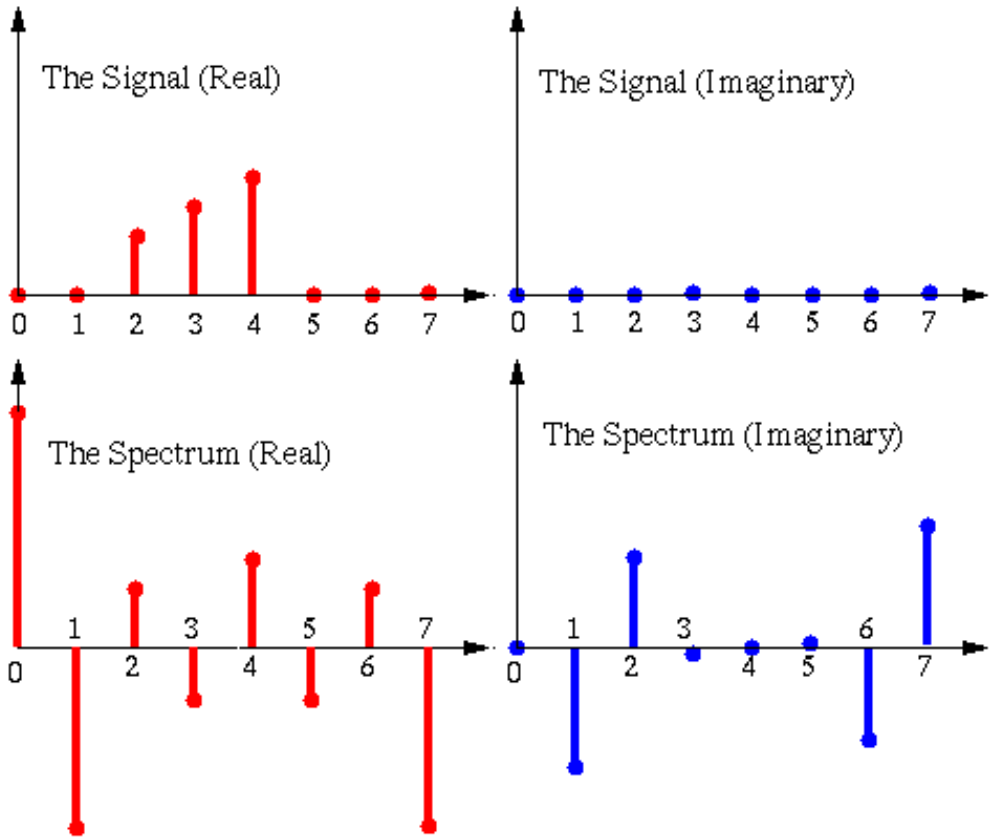
有限长度的离散时间和频率信号



离散傅里叶变换

计算

DFT 是波形信号的线性变换.



$$F[k] = \sum_{n=0}^{N-1} f[n] W_N^{kn}$$

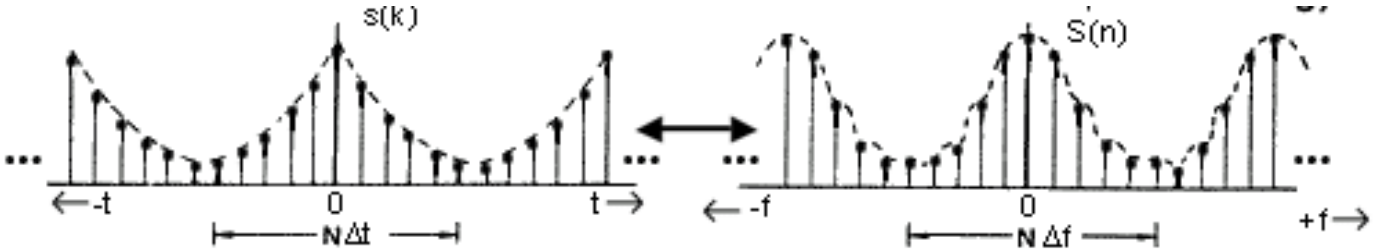
$$k = 0, 1, ...N-1$$

$$f[n] = \frac{1}{N} \sum_{k=0}^{N-1} F[k] W_N^{-kn}$$

$$n = 0, 1, ...N-1$$

$$W_N = e^{-j\frac{2\pi}{N}}$$

DFT 特性



- **对称性:**

$$|F[k]| = |F(N-k)| \quad \arg(F[k]) = -\arg(F[N-k])$$

- **能量守恒定律 (帕萨瓦尔定律 Parseval Theorem)**

$$\sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |F[k]|^2$$

DFT 中窗长的分析

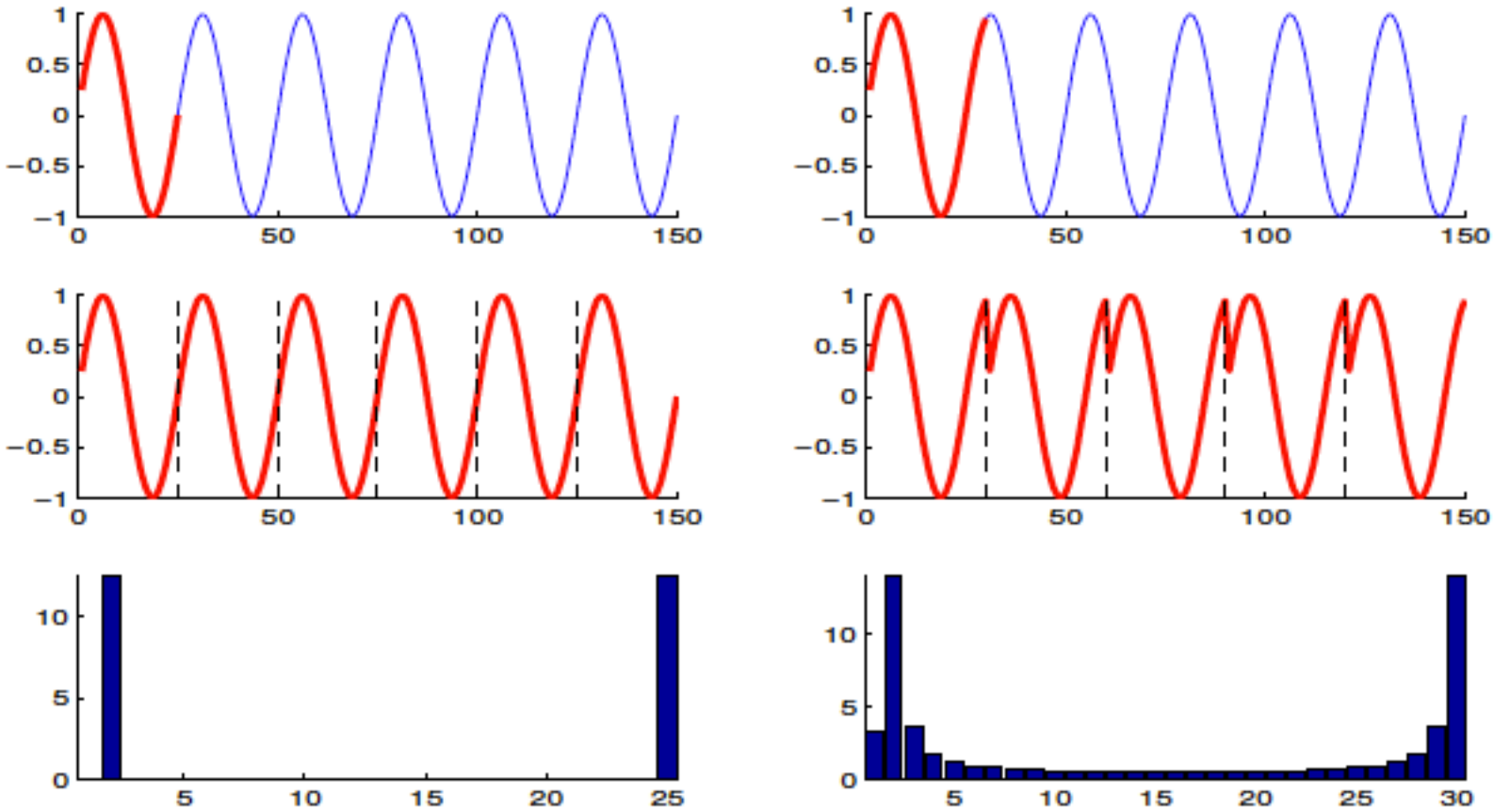
假如采样率是 $f_s = \frac{1}{\Delta T}$, 那么时域的分辨率 (resolution, duration) 是 $N\Delta T$, 频域分辨率 (resolution) 是 $\frac{f_s}{N} = \frac{1}{N\Delta T}$
当分析窗长 N 不断增加:

- **时域分辨率** 变得越差，因为突然的时域变化不能被有效的进行建模
- **频域分辨率** 变得越发准确
- **Zero-padding**: 在一个窗的语音尾部添加 0，从而增加 N
  - 可以产生更多的频点，但是不会增加真实的分辨率，因为没有新的信息被添加

DFT 潜在的周期性

- DFT 描述谱信息在 N 个频点，均匀地将谱分为了离散频点
- 仅仅周期信号才有离散频点
- DFT 在分析窗长之外，假定了周期性
- 这个假设有如下一些影响
  - 引起边界影响 (boundary/edge effect)
  - 对高频分量产生失真

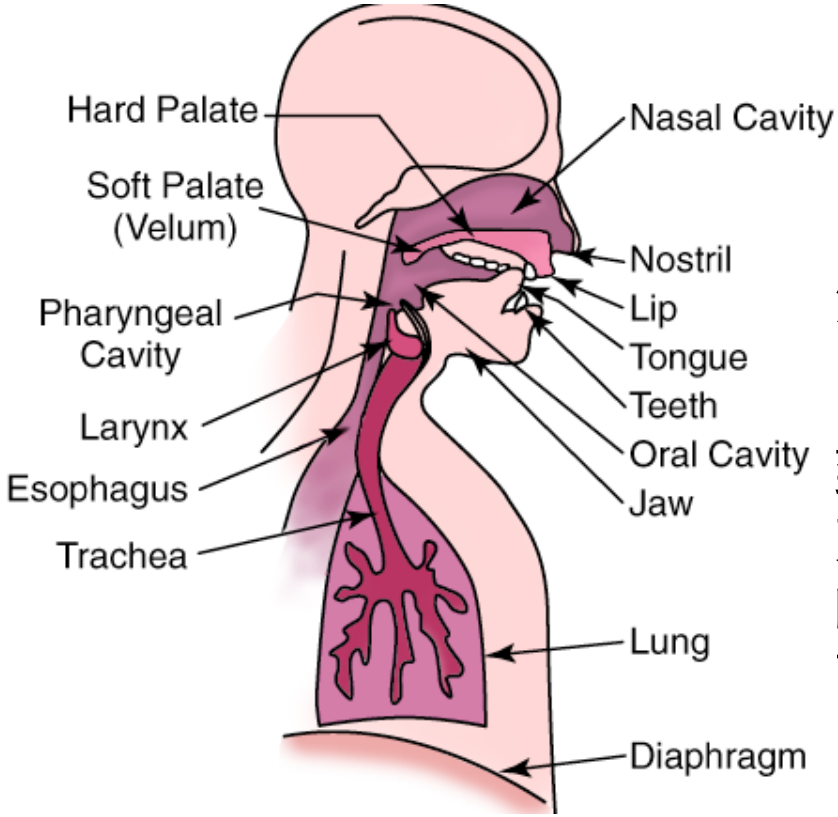
潜在周期性示例



目录

- 数字语音波形

语音波形



语音是通过人类声道中的振动活动产生的.

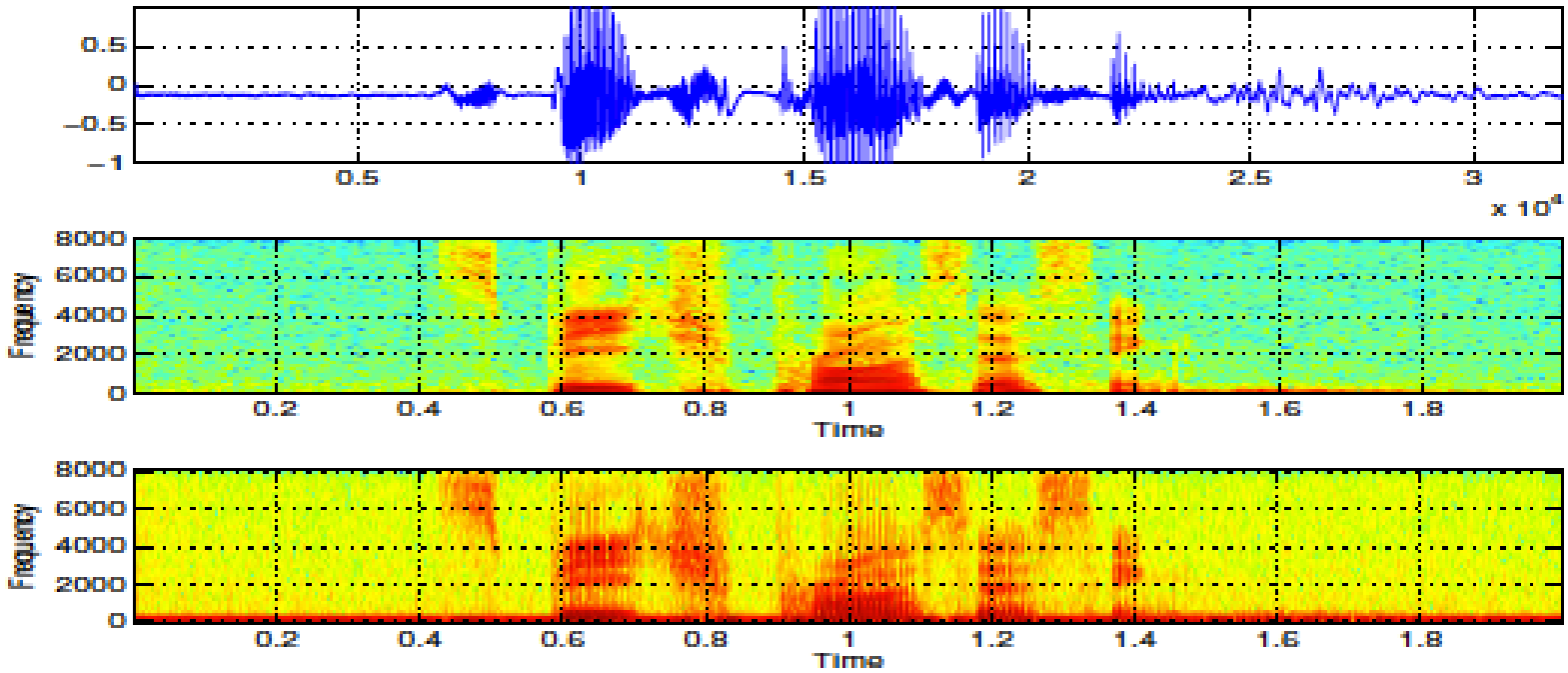
语音一般情况下通过空气传输到另一个听者的耳朵或者传输到麦克风设备。语音或者其他任何声音的都最终以空气压力变化的波形形式进行传播与扩散.

这些类型的波形都被认为是纵波。

语音信号的主要特点

语音信号带宽

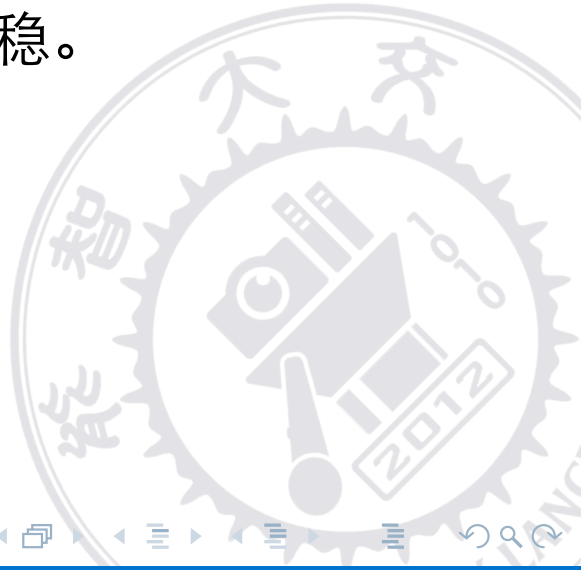
语音信号的带宽约为 5KHz，主要能量集中在低频段。下图为一段语音信号语谱图。



语音信号的主要特点

语音信号是典型的随机信号

1. 人的每次发音过程都是一个随机过程。很难得到两次完全相同的发音样本。
2. 在信号处理中，通常假设语音信号是短时平稳的。例如，可以认为在语音的浊音段部分，语音的二阶矩统计量是平稳的(在 5~10ms 内)，即二阶矩平稳，或称为宽平稳。



语音信号的时域模型

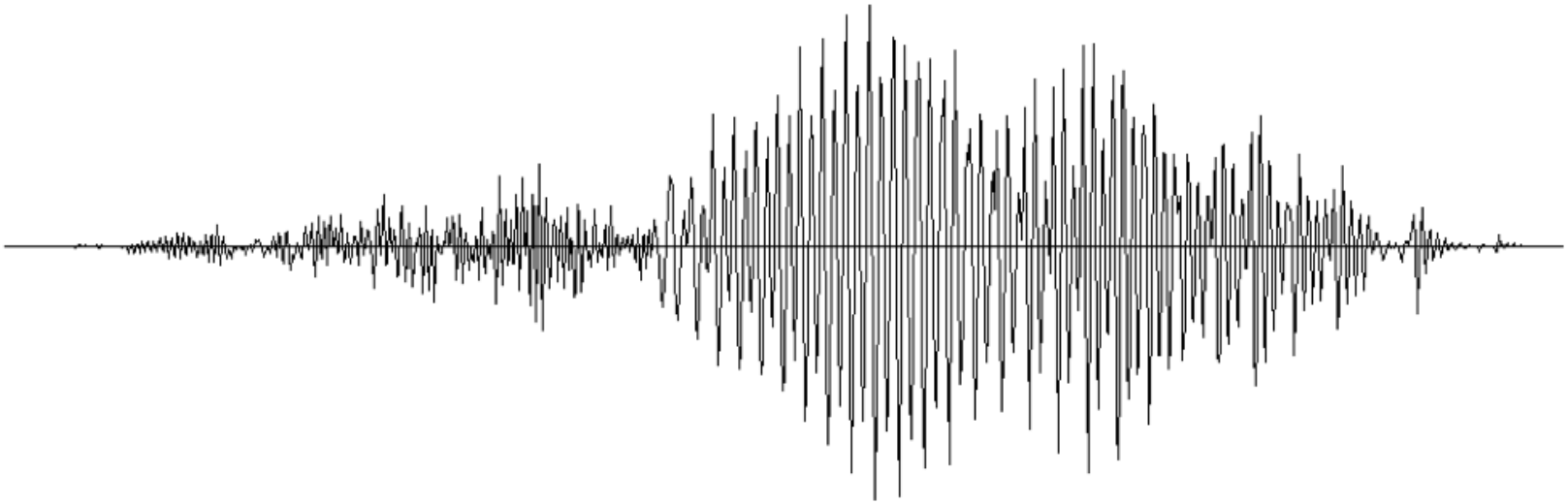


图: 语音信号的波形 (shi4)

语音信号的时域模型

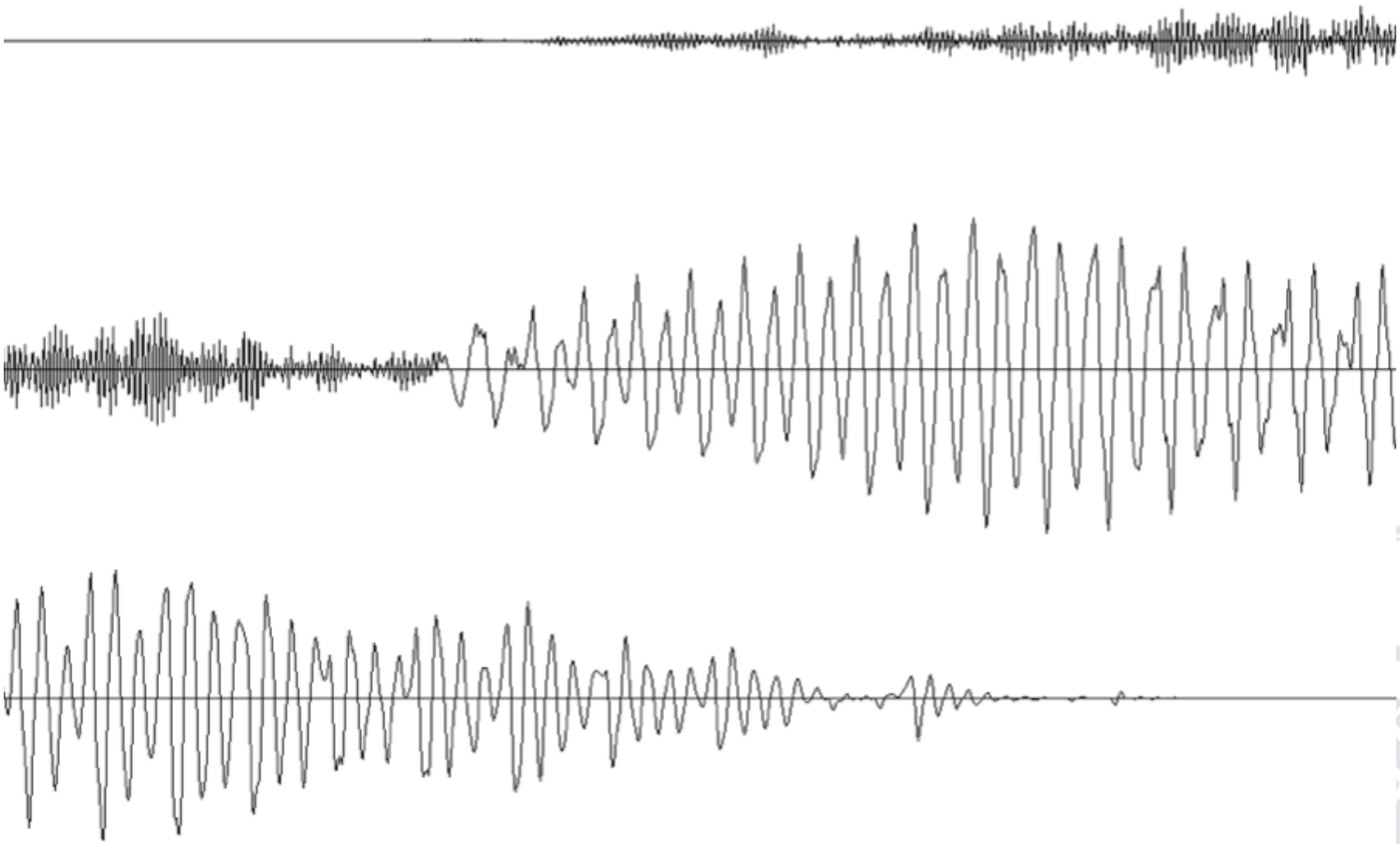


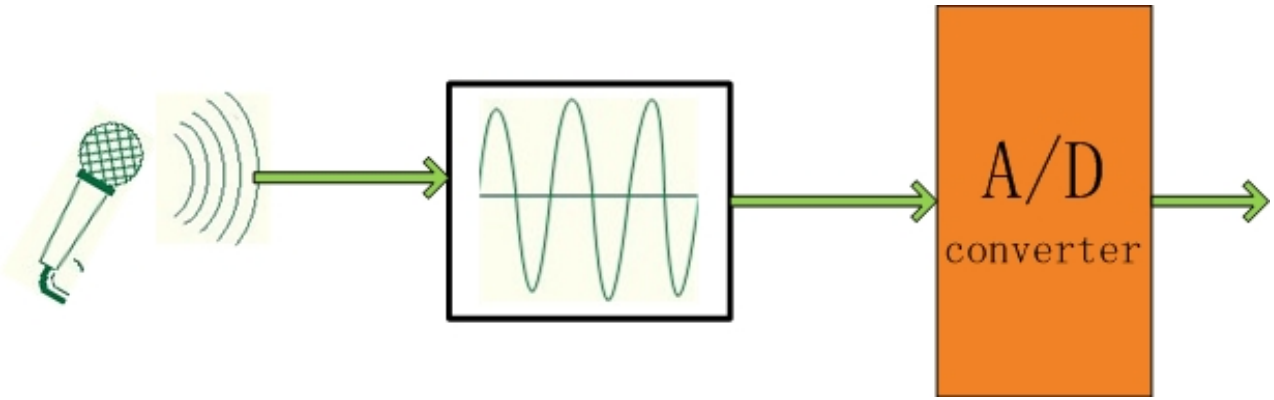
图: 语音信号的波形 (shi4) 的局部细节

语音时域信号特征

语音时域信号的特点

1. 清音段: 能量低, 过零率高, 波形特点有点像随机的噪声。这部分信号常与语音的辅音段对应。
2. 浊音段: 能量高, 过零率低, 波形具有周期性特点。所谓的短时平稳性质就是处于这个语音浊音 (元音) 段中。
3. 过渡段: 一般是指从辅音段向元音段信号变化之间的部分。信号变化快, 是语音信号处理中最复杂、困难的部分。

数字语音波形

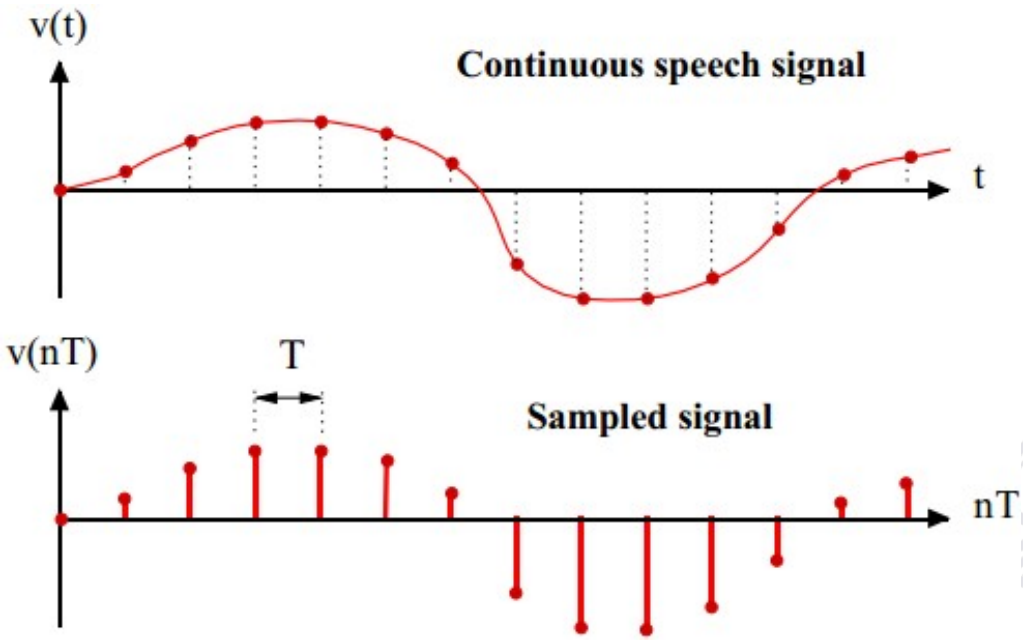


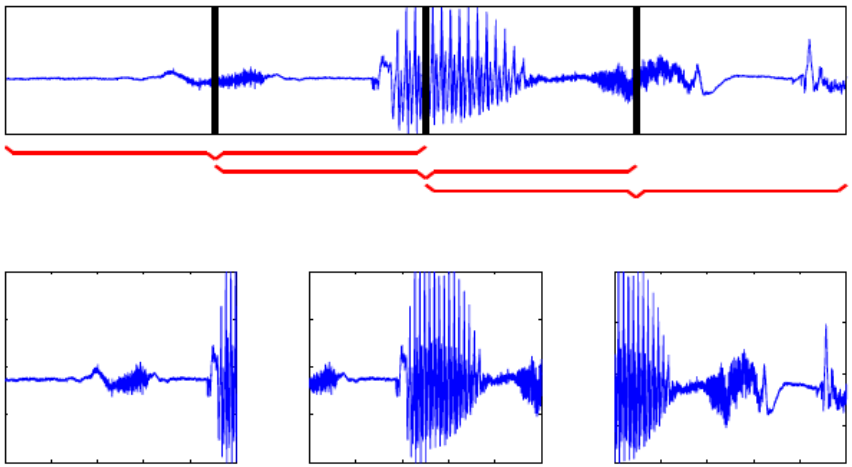
当我们对着一个麦克风说话的时候, 声压的变化被转化成电压层面的成比例的变化。一台装配有合适硬件的计算机通过一个被称为模数转换 (ADC) 的过程，可以将模拟电压信号变化转化成数字声音的波形信号，过程包括:

数字语音波形 (Cont'd)

- **采样:** 在相等的时间间隔，对连续变化的语音信号进行采值，得到每一个固定的声压数值. 通常情况下是采样率是 16,000 次或者 8,000 次每秒, i.e. 16K Hz for PC 机，8k for 电话.
- **量化:** 将每一个采样的波形幅值表示成离散的数值 (在一个给定量化 bit 数量的可表示数量范围内，近似成最相近的数值). 例如, 8 bits 和 16 bits 可以分别表示总共 256 或者 65536 个可能的量化层次.
- **压缩:** 将已经量化的数值表示成更加紧凑的形式，从而节省传输或者存储时的空间。通常的压缩形式包括线性 PCM, μ-law, A-law 等.
- 数字语音波形是**一个随机离散信号**

通过采样，原来模拟的" 连续时间" 信号可以转换成数字的" 离散时间" 信号.

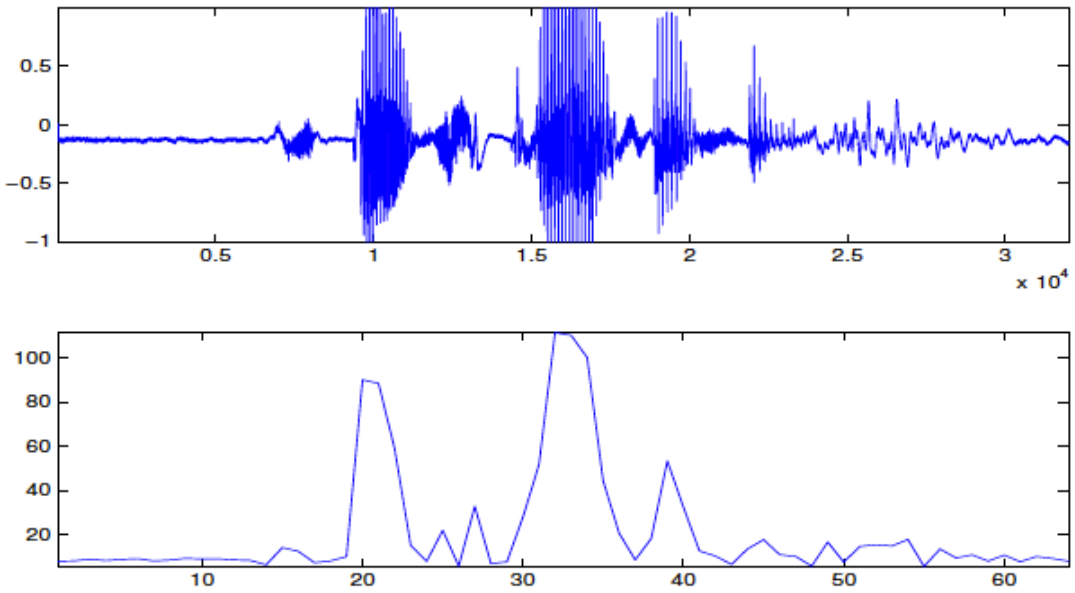




- 完整语音波形是一个很长非平稳长时采样序列. 在语音信号处理中，将长时序列分成准平稳的不同的块/帧是很有用的
- **帧大小**: 选取时应该考虑如下因素的折中方案:
  - 有足够多的采样点来用于准确的语音信号特性的分析
  - 确保准平稳假设的有效性
- **帧移**: 前后两帧语音相互重叠的采样点部分. 使用部分重叠的帧选取方法可以更好地表示信号的动态特性

基本波形处理

短时 (帧) 能量



短时能量: 一帧内的采样点的平方和:

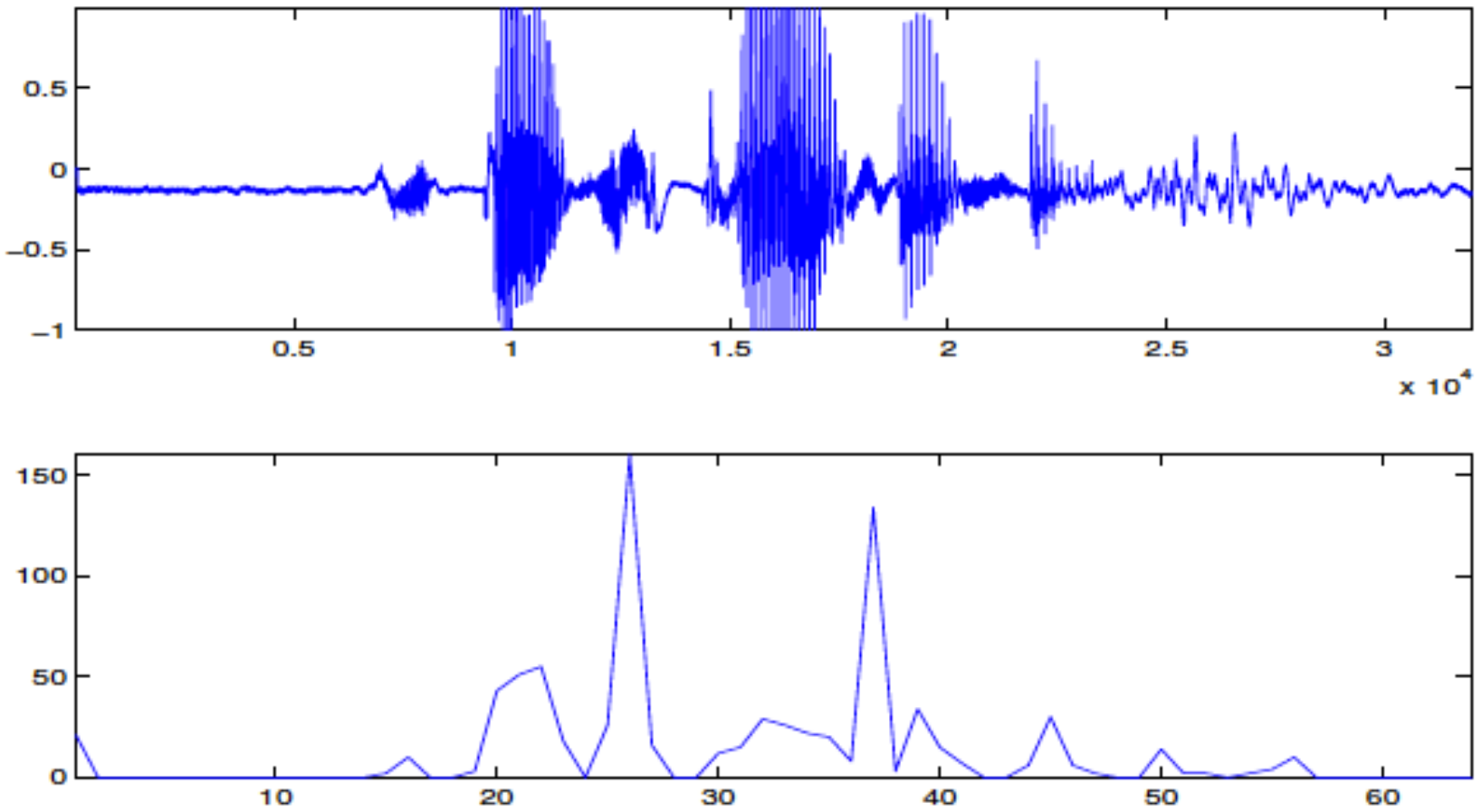
$$E = \sum_{i=0}^{N-1} s_i^2$$

过零率 Zero-crossing rate (ZCR): 每一帧中采样点正负反复的次数 (跨越零点的次数). 清音的过零率较大，浊音的过零率较小.

```
int ZCR(float s[]) {
  int count = 0;
  for (int i=1; i<s.length; i++)
    if (s[i-1]*s[i] <= 0)
      count++;
  return count;
}
```

基本波形处理

过零率 (II)



1. 短时能量:

$$E = \sum_{n=0}^{N-1} s^2(n) \tag{1}$$

2. 短时平均振幅:

$$M = \sum_{n=0}^{N-1} |s(n)| \tag{2}$$

3. 短时过零率:

$$Z = \frac{1}{2} \left\{ \sum_{n=0}^{N-1} |sgn[s(n)] - sgn[s(n-1)]| \right\} \tag{3}$$

$$其中 sgn[n] = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

对于能量受限的时域信号 $f(t)$, 它的傅里叶变换可以写成

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t)\mathrm{e}^{-j\omega t}\mathrm{d}t \tag{4}$$

以上这个傅立叶变换, 在“宏观上”给出信号 $f(t)$ 的频谱信息, 但是却无法确定某个“局部”时间段频谱的确切信息。

如果谱分析不能确定这种时间序列的次序 (即位置), 那么这种信号分析的手段在应用上就会受到限制。同时我们也希望能够通过观测到的局部时域信号的频谱信息来了解 (构造) 整个 $f(t)$ 的频谱信息。

有许多技术都可以用来完成信号的短时谱分析。最典型的就是小波变换和我们现在常采用的傅立叶短时谱分析技术。下面我们来研究短时傅立叶谱分析技术。

傅立叶短时谱分析与窗的形状和位置有关 (与时刻有关)。假设窗函数为 $w(t)$, 那么信号 $f(t)$ 的短时傅立叶变换为

$$\hat{f}_w(\omega)|_{t_0} = \int_{-\infty}^{\infty} f(t)w(t-t_0)\mathrm{e}^{-j\omega t}\mathrm{d}t \tag{5}$$

例如, 如果选择窗的形式为一个高斯函数 $w(t) = \frac{1}{2\sqrt{\pi a}}\mathrm{e}^{-\frac{t^2}{4a}}$, 这个窗口函数有如下性质

$$\int_{-\infty}^{\infty} w(t-t_0)\mathrm{d}t_0 = \int_{\infty}^{\infty} w(t)\mathrm{d}t = 1$$

所以有

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}_w(\omega)|_{t_0} \mathrm{d}t_0 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t)w(t-t_0)\mathrm{e}^{-j\omega t}\mathrm{d}t\mathrm{d}t_0 \\ &= \int_{-\infty}^{\infty} f(t)\mathrm{e}^{-j\omega t}\int_{-\infty}^{\infty} w(t-t_0)\mathrm{d}t_0\mathrm{d}t \\ &= \hat{f}(\omega)\end{aligned} \tag{6}$$

这说明 $\hat{f}(\omega)$ 可以被加窗后的短时谱 $\hat{f}_w(\omega)|_{t_0}$ 所精确地分解。这正是我们所希望的性质。

更一般地, 若 $X(r,\omega)$ 是语音序列 $x[n]$ 的在时刻 r 的短时傅利叶变换

$$X(r,\omega)=\sum_{m=-\infty}^{\infty}x[m]w[r-m]\mathrm{e}^{-j\omega m} \tag{7}$$

若满足条件

$$\sum_{r\in S}h[n-r]w[r-n]=1,\forall n\in Z,S\text{为短时谱取样时刻值的集合} \tag{8}$$

则语音序列 $x[n]$ 可以由短时谱精确重构

$$x[n]=\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_{r\in S}h[n-r]X(r,\omega)\right]\mathrm{e}^{j\omega n}\mathrm{d}\omega \tag{9}$$

其中 $\sum_{r\in S}h[n-r]X(r,\omega)$ 项可以理解为利用插值滤波器 $h[r]$ 得到在 n 时刻的短时谱。

4. 对于时域离散信号 $x(n)_{\infty}$, 短时傅立叶变换定义

$$X_n(\mathrm{e}^{j\omega}) = \sum_{m=-\infty} x(m)w(n-m)\mathrm{e}^{-j\omega m} \tag{10}$$

这里 $w(n)$ 为窗函数。例如, 常用的窗函数有

- 矩形窗:

$$w(n) = \begin{cases} 1, & 0 < n < N-1 \\ 0, & \text{其他} \end{cases}$$

- 汉明窗:

$$w(n) = \begin{cases} 0.54 - 0.46\cos(2\pi n/(N-1)), & 0 < n < N-1 \\ 0, & \text{其他} \end{cases}$$

- 汉宁窗 (Hann):

$$w(n) = \frac{1}{2}\left[1 - \cos(\frac{2\pi n}{N-1})\right], 0 < n < N-1$$

- 巴特利特窗 (Bartlett)(三角形窗):

$$w(n) = \begin{cases} \frac{2n}{N-1}, & 0 < n < \frac{N-1}{2} \\ 2 - \frac{2n}{N-1}, & \frac{N-1}{2} < n < N-1 \end{cases}$$

- 布莱克曼窗 (Blackman):

$$w(n) = 0.42 - 0.5\cos(\frac{2\pi n}{N-1}) + 0.08\cos(\frac{4\pi n}{N-1}), 0 < n < N-1$$

基本波形处理

窗函数性质

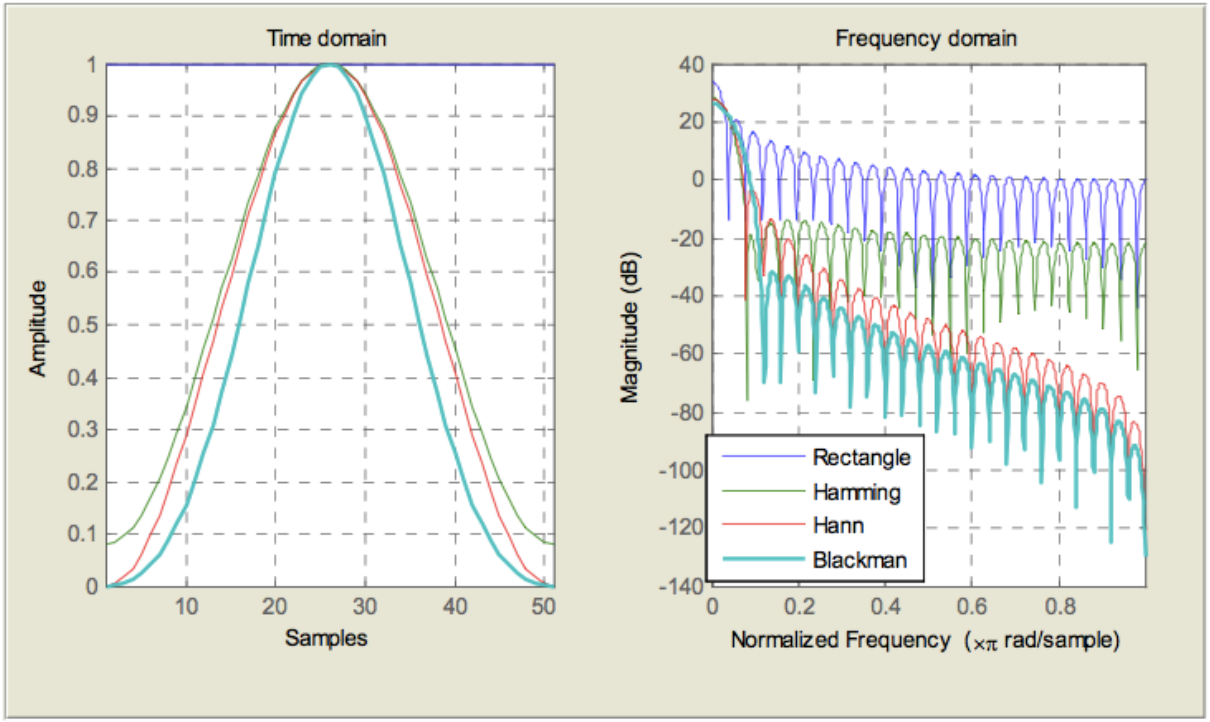
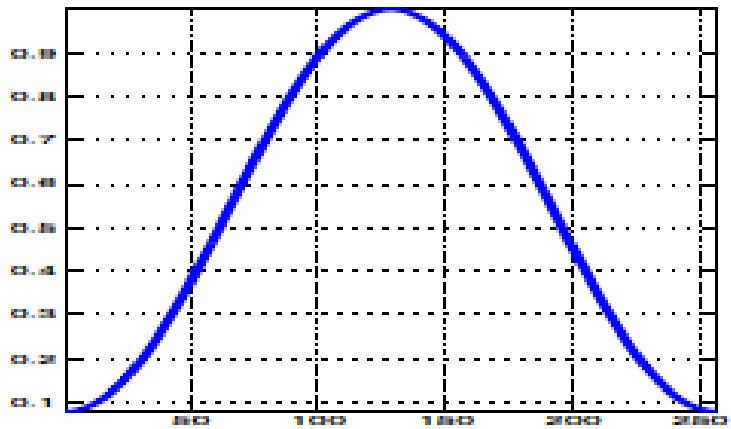
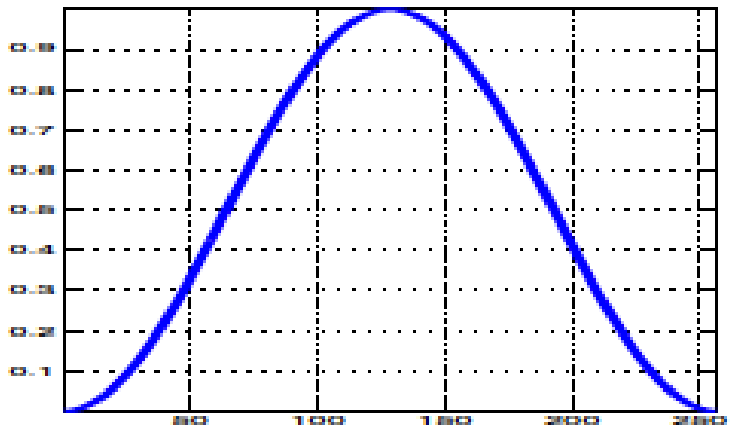


图: 各种窗函数时域频域特性比较

加窗





在信号处理中, 窗函数 是一个在所选择的区域外都是零值的函数. 用窗函数去乘信号，从而减少或降低窗边界的影响.

- **Hamming 窗:**
  $w(n) = 0.53836 - 0.46164\cos(\frac{2\pi n}{N-1}),\ \ n = 0, \cdots, N-1$
- **Hanning 窗:** $w(n) = 0.5\left(1 - \cos(\frac{2\pi n}{N-1})\right)$

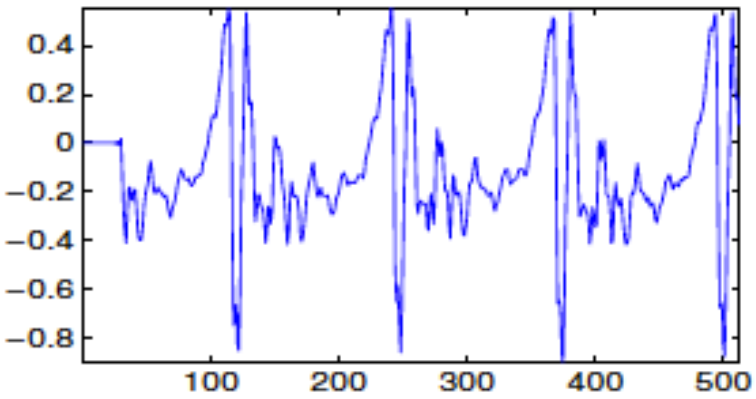
- **动机:** 在一个频段范围内, 用来增加某一些频率的幅值 (一般是较高的频率) w.r.t. 降低另一些频率的幅值 (一般是较低的频率), 从而来增加信号整体的信噪比.
- **计算:**

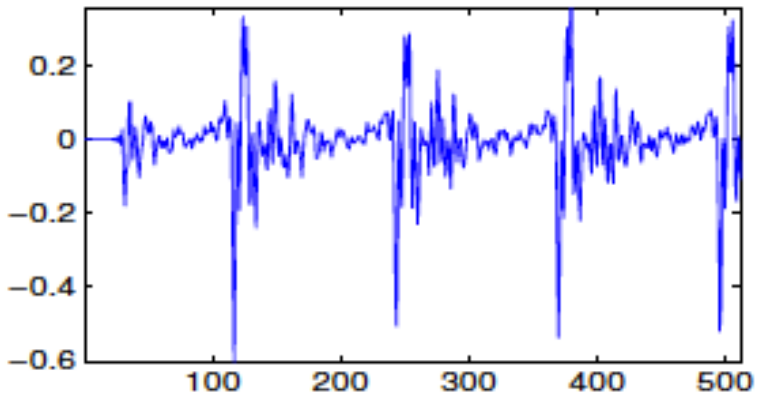
$$x_n = x_n - \alpha x_{n-1}$$

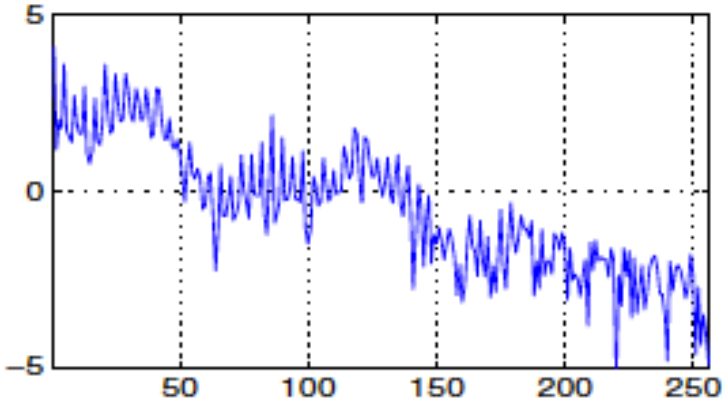
其中 α 是预加重系数 (典型值取 0.97). 边界情况假定 $x_{-1} = x_0$.

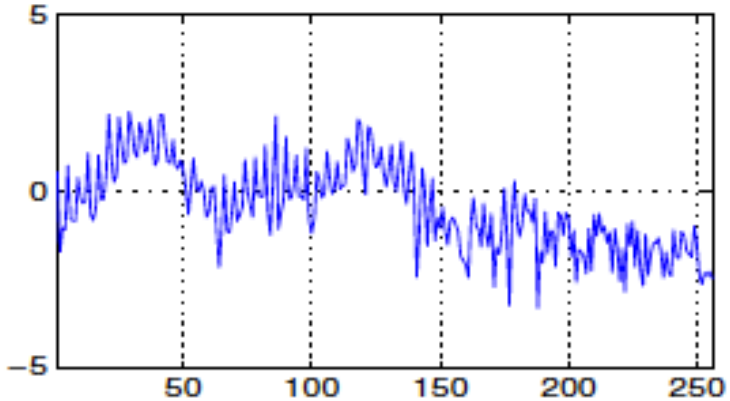
基本波形处理

有无预加重的谱分析









自相关 强调的是语音波形的周期性. 正常情况下在一个较宽的窗下进行计算.

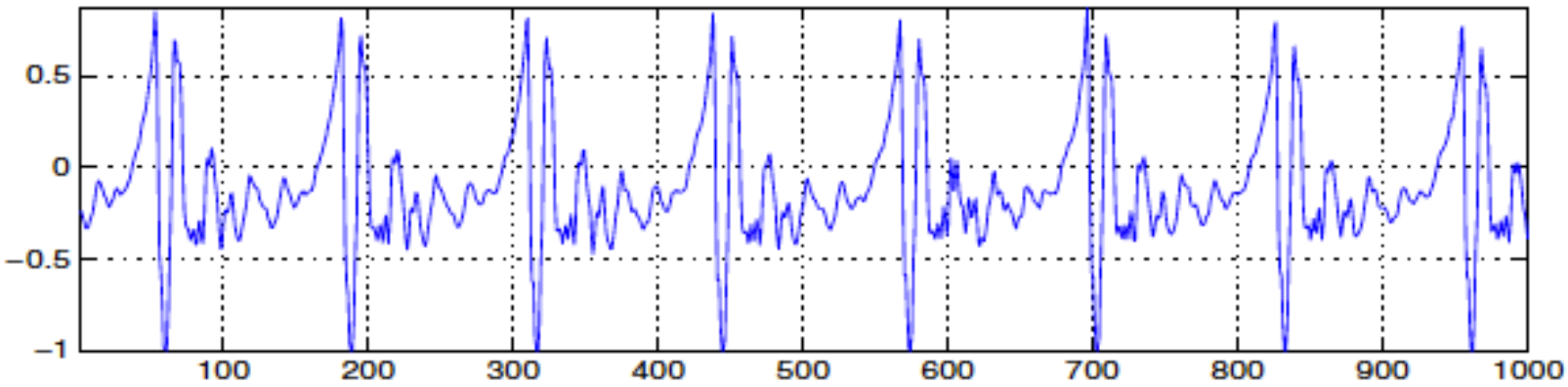
$$r_k = \sum_{i=0}^{N-k-1} s_i s_{i+k}$$

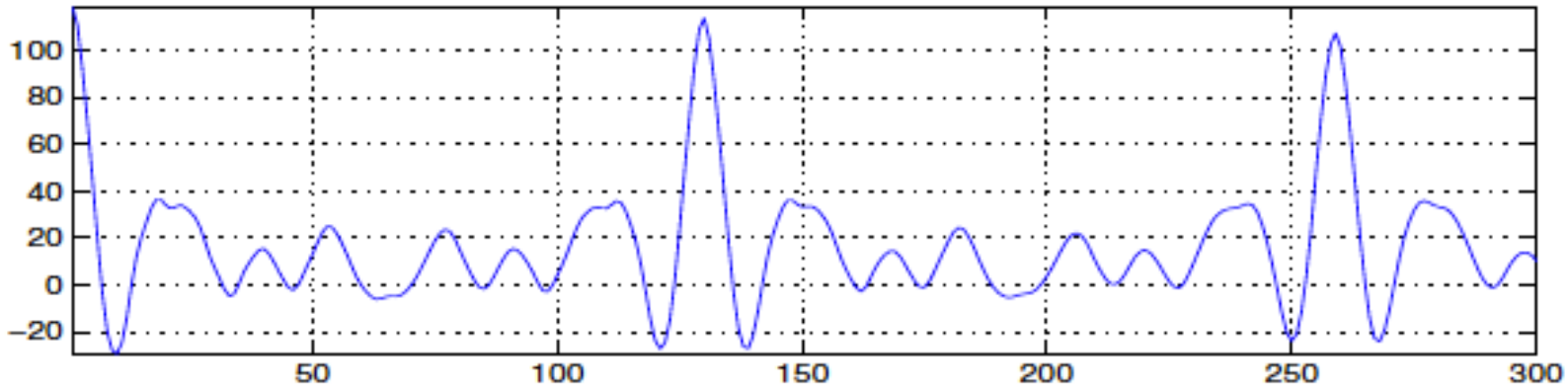
其中 k 是相关周期.

```
float AutoCorr(float s[], int k) {
  float sum = 0;
  for (int i=0; i<s.length-k; i++)
    sum += s[i] * s[i+k];
  return count;
}
```

基本波形处理

自相关


0.5
0
−0.5
−1
100
200
300
400
500
600
700
800
900
1000


100
80
60
40
20
0
−20
50
100
150
200
250
300

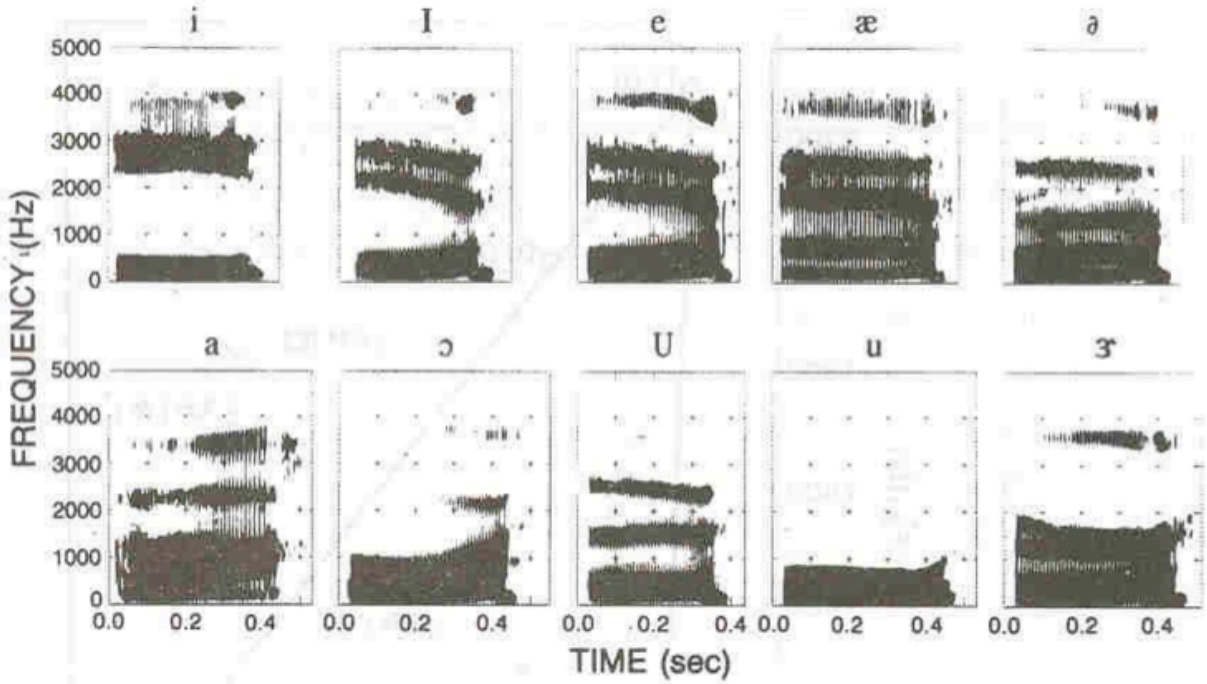
目录

- 谱分析

语音信号的短时谱特征

语谱图

横轴表示时间，纵轴表示频率，用灰度表示对应频谱分量的信号强度。



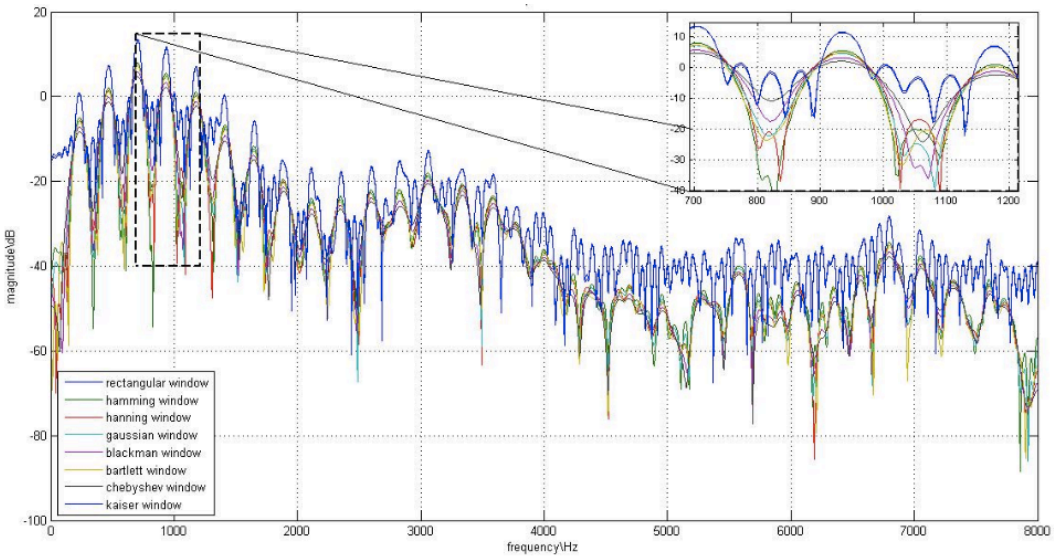
Spectrograms of the vowel sounds.

浊音谱特征

浊音谱的谱线结构

谱线结构是与浊音信号中的周期信号密切相关的。具有与基音及其谐波相对应的谱线。

清音的频谱无明显的规律, 比较平坦。在语音识别中使用统计模型的方法加以解决。



浊音谱特征

浊音谱的共振峰结构

频谱包络中有几个凸起点，与声道的谐振频率相对应。这些凸起点称为共振峰 (Formant)。其频率称为共振峰频率。按频率由低到高依次为第一共振峰、第二共振峰... 。相应频率用 F1、F2、F3... 来表示。

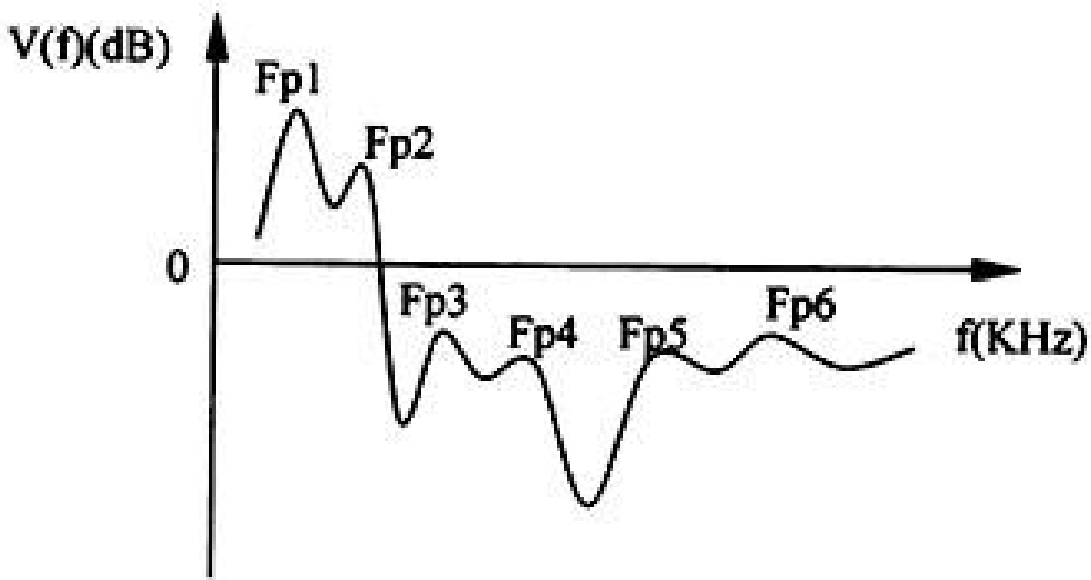


图1 声道频域特性（频率响应图）

浊音谱特征

浊音谱的共振峰结构

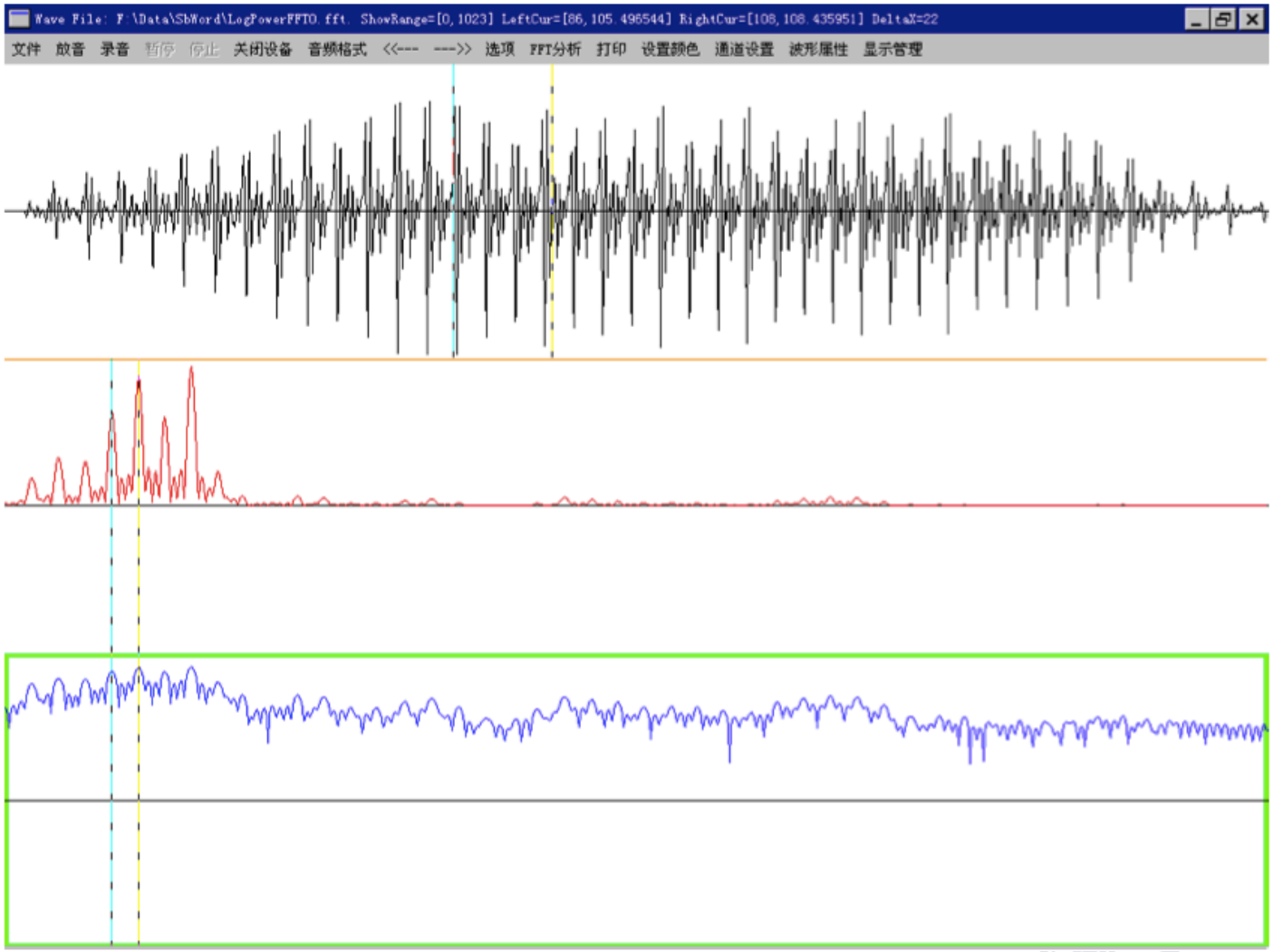
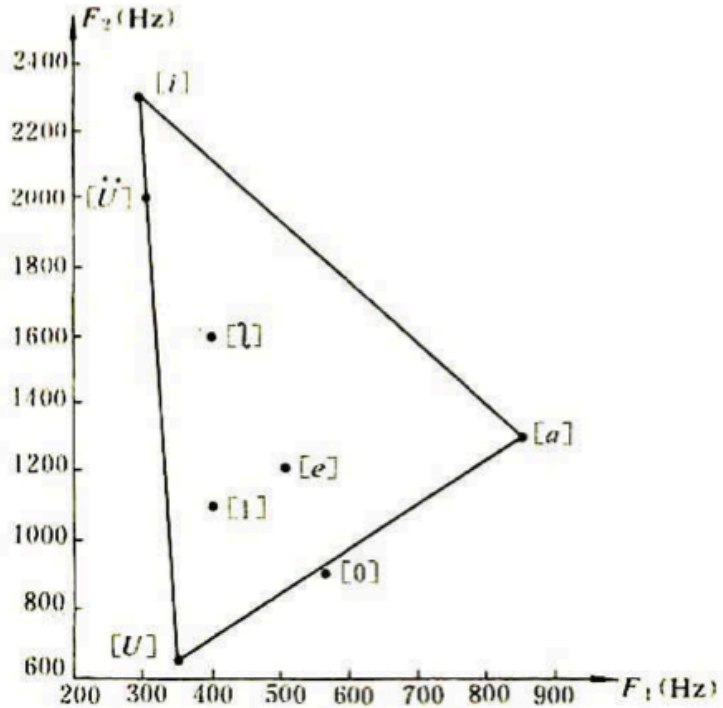


图: 浊音信号的傅立叶分析谱

浊音谱特征

元音三角形图

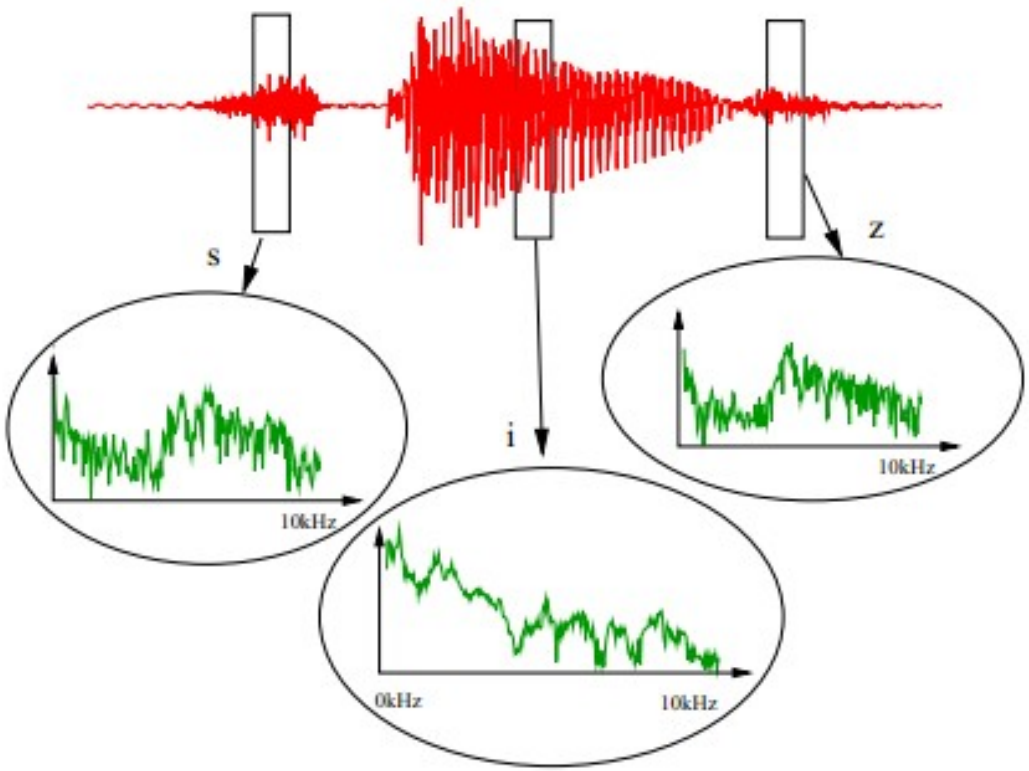
所谓的元音三角形图就是指不同元音的 F1、F2 共振峰频率在平面图上的关系。



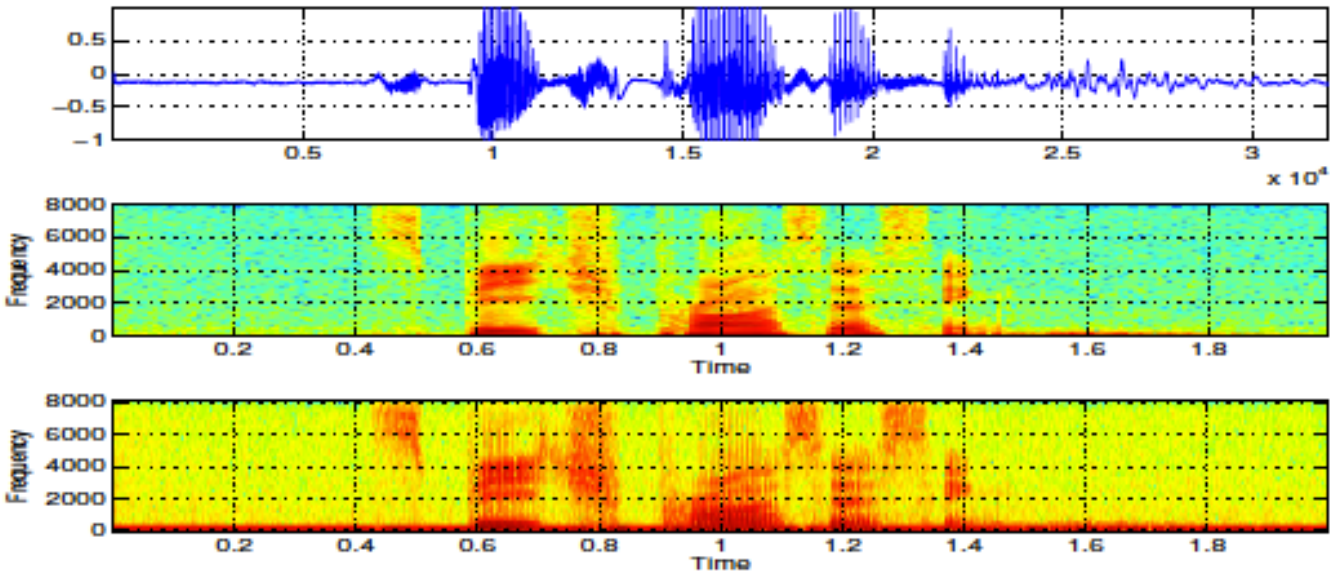
元音三角形

语音信号处理 - 短时傅里叶变换

- 假定语音信号是一个准平稳过程
- 将语音信号切分为短时的片段, e.g. 10ms
- 针对每一个片段应用离散傅里叶变换 DFT



语谱图

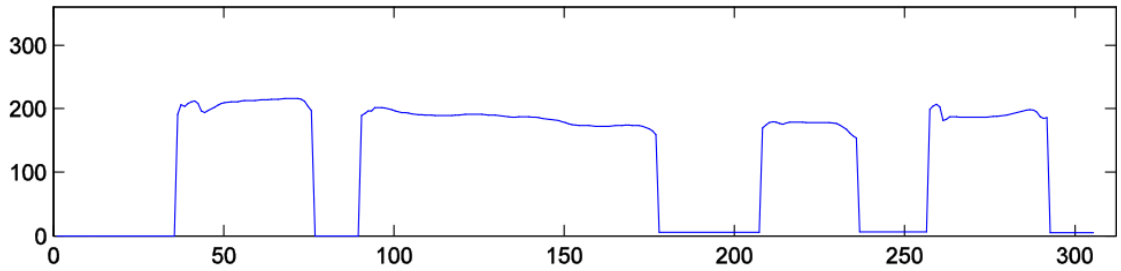


- 时域和频域的分辨率不得不做一个折中选择
- 调整**窗长** 和**窗之间的帧叠**，从而来做两者之间的折中
- *Middle:* 256 点的窗长，50% 的窗叠 (较好的频域分辨率)
- *Bottom:* 64 点的窗长，50% 的窗叠 (较好的时域分辨率)

1. 基音的周期就是声带振动的周期。基音周期的倒数就是基音频率。
2. 基音是与人的声带长度、质量等物理量有关。因此与人的年龄、性别、情绪等生理状态有关。

[注意]: 音高 (Pitch) 与基音的关系。音高是听觉量, 基音是物理量。正如冷热与温度的关系一样。

基频 (基本频率)



基频 (Pitch) or **基本频率 (fundamental frequency)** F_0 是人类语音的最低的频率，它是组成语音中更高频率成分的基础.

- 浊音部分有明显的周期性，而清音部分仅仅是随机振动
- 人类的 F_0 范围大概范围是 60Hz 到 300Hz

基音与四声

基音周期与基音频率

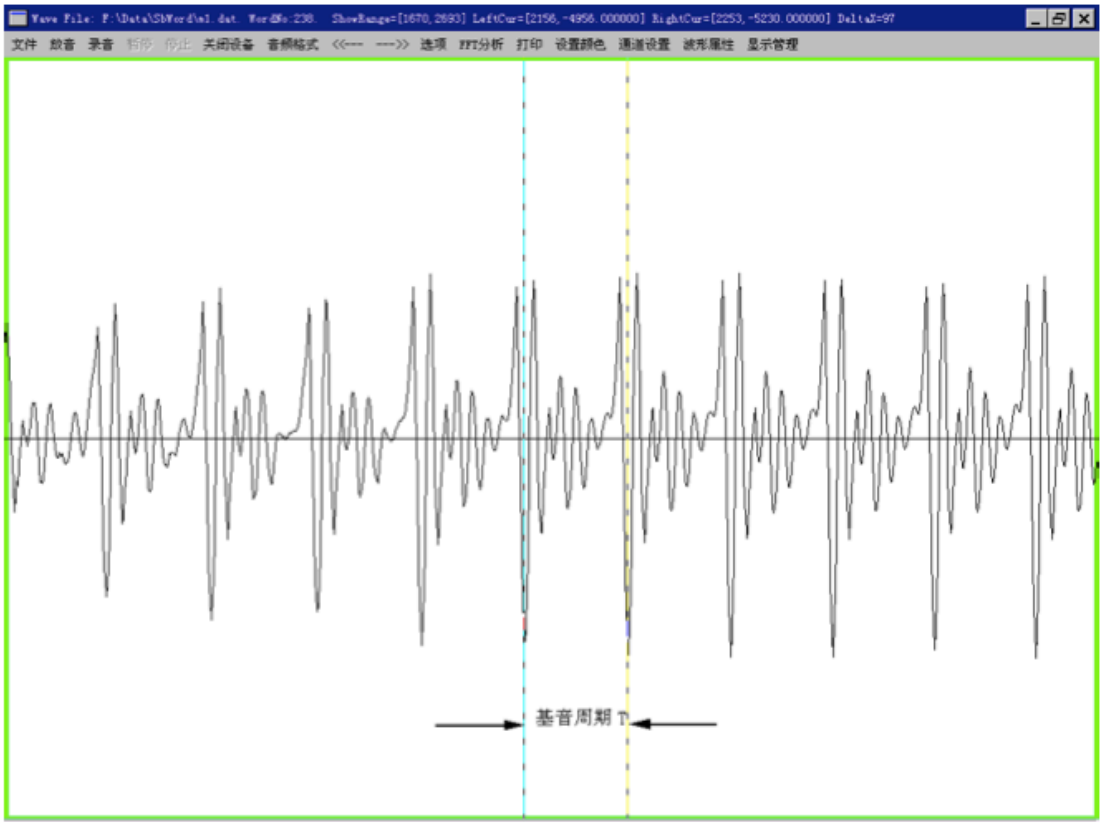


图: 基音周期示意图

1. AMDF 法
   定义平均幅度差函数

$$\gamma(l) = \sum_{n=0}^{N-l-1} |S_w(n+l) - S_w(n)|$$

在这里 $S_w(n)$ 是加窗截取的一段语音信号。
假设 T 为语音信号的基音周期, 当 $l = nT$, $n = 1, 2, \cdots$ 时, $\gamma(l)$ 函数接近局部极小值。
AMDF 算法特点: 只用到简单的加减法运算, 没有使用乘法运算。适合于早期普通的 CPU, 因为这种 CPU 的乘法操作要比加减法操作费时。

2. 自相关法
   定义语音的自相关函数

$$R(l) = \sum_{n=0}^{N-l-1} S_w(n+l) S_w(n)$$

当 $l = nT$, $n = 1, 2, \cdots$ 时, $R(l)$ 函数接近局部极大值。
自相关法特点: 在这个算法中使用了乘-累加操作。在数字信号处理器中有专门的硬件指令来快速完成 (只要一个周期) 这种乘-累加运算。因此这个算法在 DSP 中得到了普遍的应用。

1. 无论是使用 AMDF 法或是自相关法求语音信号的基音周期, 都要在基音周期 T 的范围内 $[T_{min}, T_{max}]$。

2. 搜索 $\gamma(l)$ 或 $R(l)$ 的极值点位置。

3. 一般取 $0.5T_{min} < l < 1.5T_{max}$ ，先计算所有的 $\gamma(l)$ 或 $R(l)$ 值, 然后再搜索得到基音。

基音的检测

时域上的基音检测方法

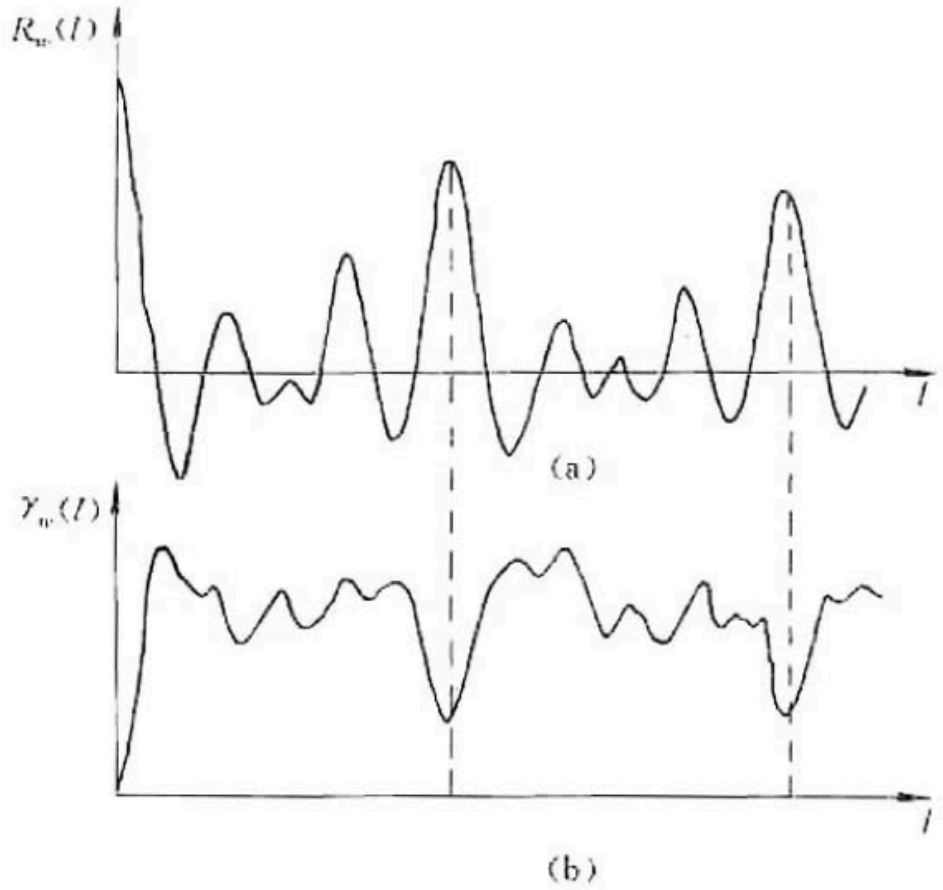
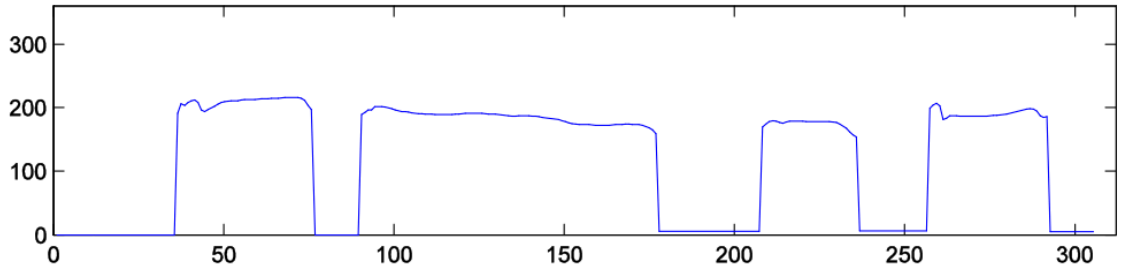


图: 语音 (浊音) 的自相关函数和 AMDF 曲线

基频的检测

时域上的基音检测方法



1. 计算自相关 r_0 到 r_{max}，其中 r_0 表示的是能量
2. 在 r_0 到 r_{max} 的范围内找到峰值 r_p
3. 假如 $r_p > 0.3r_0$ ，说明此语音是周期为 p 的浊音
4. 否则，这一帧语音是清音

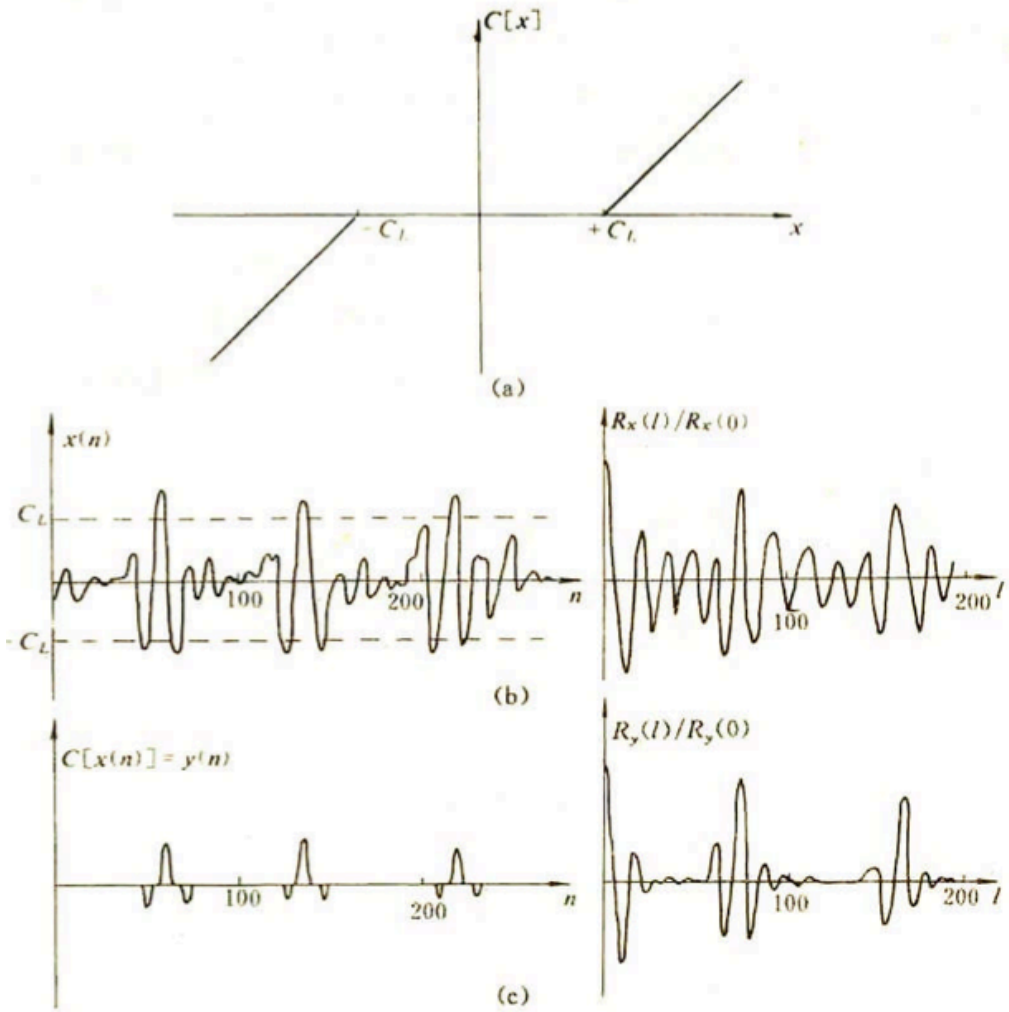
3. 中心削波法

在计算语音信号的自相关函数时，为了提高效率，减少干扰，可以先对语音信号进行中心削波，然后再计算自相关函数。

根据实验观察，自相关函数 $R(l)$ 的局部峰值点位置与语音幅度的峰值点位置重合。根据这个特点，在自相关法中只需要计算这些峰值点位置的自相关函数 $R(l)$，然后再搜索比较即可得到信号的基音周期。

基音的检测

时域上的基音检测方法

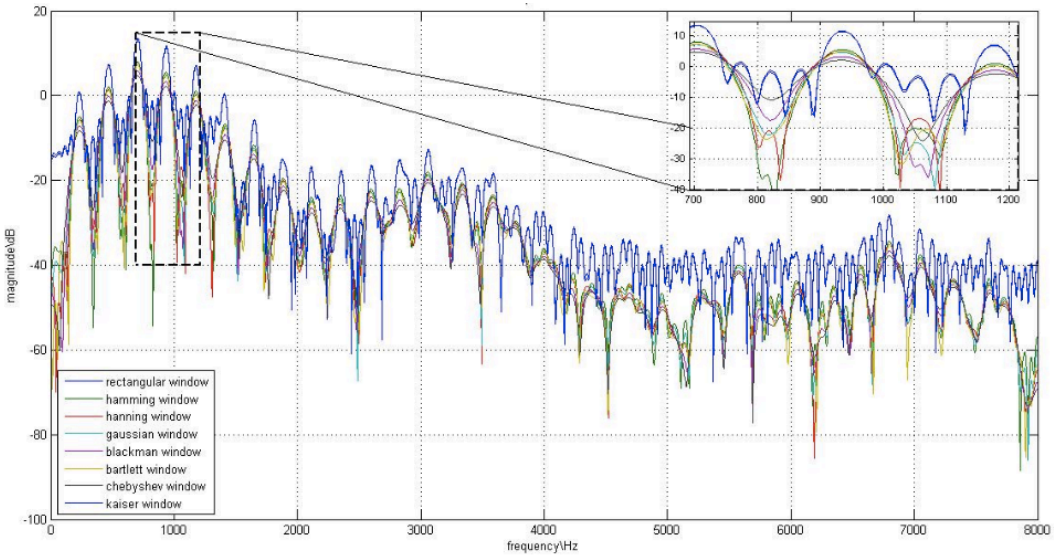


语音信号经过中心削波后自关函数具有更尖锐锋起的示例

基音的检测

频域上的基音检测方法

在频域中，常常是用谐波分析法，即对浊音信号的谱线结构进行分析来计算得到基音周期。



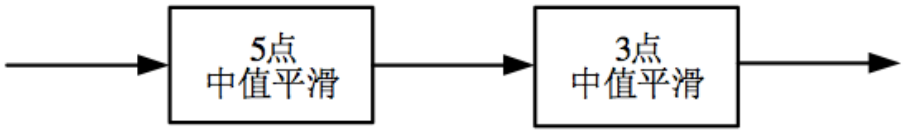
1. 由于在基音的提取过程中不可避免地要产生误差, 主要是基音周期减半或加倍的现象 (根据方法的不同, 误差的现象会有所不同)。

2. 一般情况下 90% 左右的基音周期都会被准确提取, 但是总有少部分的基音是提取不准确的。因此需要采取平滑的方法去掉这些奇异点。

3. 在语音编码和汉语四声识别中, 基音平滑直接影响到系统的性能。

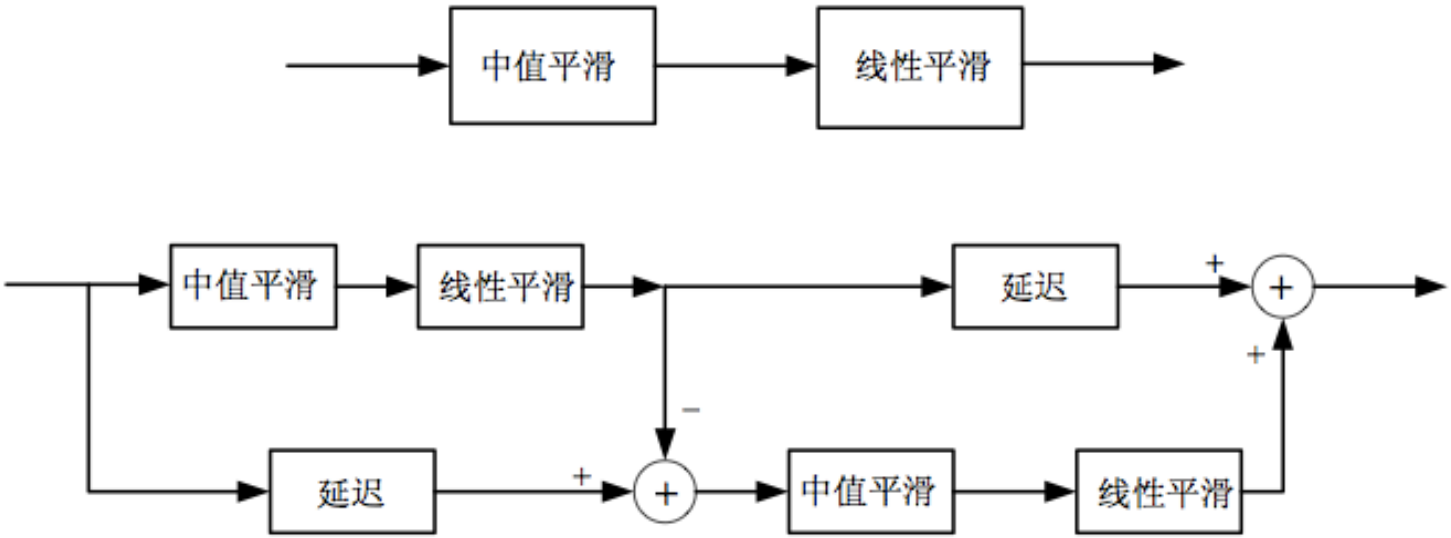
几种常用基音平滑方法

1. 非线性平滑
   例如: 采用中值平滑



2. 线性平滑
   例如: 采用 FIR 滤波器进行低通滤波平滑

3. 组合平滑
   例如: 前两种方法的组合



汉语中的声调, 对汉语的系统 (识别, 合成等) 都很重要。
汉语的声调起着辨字、辨义的作用。

- 阴平———一声
- 阳平———二声
- 上声———三声
- 去声———四声

对于孤立字音节的声调轨迹, 一般可以分成三段

1. 弯头段: 对应于音节发音开始时的过渡段。
2. 调型段: 对应于音节的饱满发音过程。
3. 降尾段: 对应于音节结束时的过渡段。

调型段在汉语四声识别中起主要作用。

汉语孤立字的基音调式和四声

汉语四声与基音频率的关系

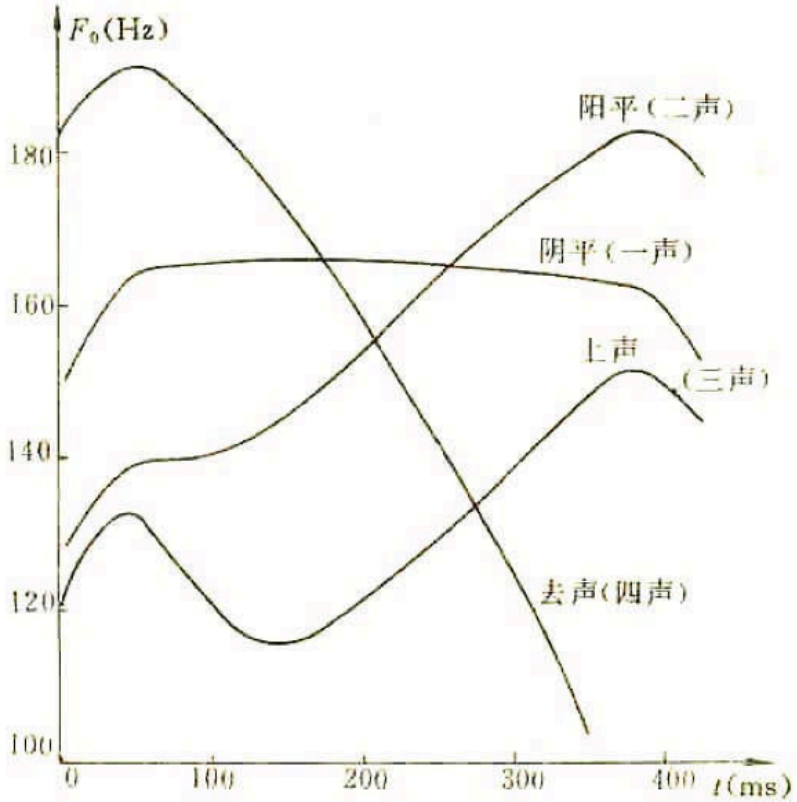
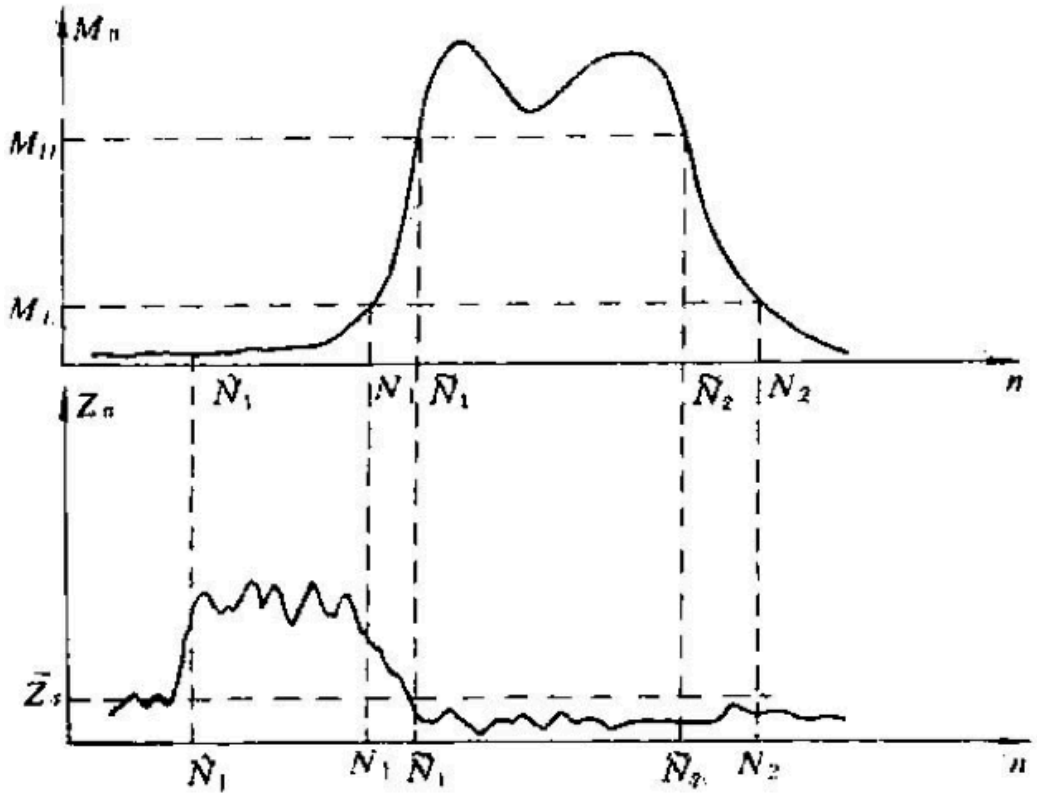


图: 汉语四声与基音轨迹示意图

[注]

1. 一声的平均基音频率要高于三声的平均基音频率，一般来说三声的平均基音频率是最小的。二声和四声较容易区分。
2. 在孤立字语音中，这种调式与基音的轨迹一一对应。但是，在连续语音中，基音与调式无明确固定的对应关系。容易因为受到协同发音的影响，调式变得更加复杂，需要进行特殊处理。
3. 基音的估计对谱分析，特别是对语音合成编码起着决定性的重要作用。

在实验室较为安静的环境下，利用短时能量和过零率特征可以得到较为满意的语音端点检测结果。更进一步地，通过判断在语音中是否存在合理的基频值，可以过滤掉绝大部分的非语音干扰。与课程 Project-1 中，第一部分通过语音信号分析的方法进行语音端点检测相关。



Project-1

语音端点检测：

- 基于线性分类器和语音短时能量的简单语音端点检测算法
  - 利用语音的短时特征 (短时频谱，短时能量，过零率，甚至基频等)
  - 简单状态机的使用
  - 简单线性分类器的使用 (如：阈值分类器)
  - Technical Report, for all the students.
- 基于统计模型分类器和语音频域特征的语音端点检测算法
  - 利用语音的频域特征 (MFCC,PLP, FBank 等)
  - 简单状态机的使用
  - 统计模型分类器的使用 (GMM 或者 NN)
  - Technical Report, for all the students.

目录

- 听觉系统与听觉特性

听觉系统

研究和理解基本的听觉系统与听觉特性, 对设计性能优越的语音系统 (识别, 合成等) 也非常重要。

一个完整的语音通信系统总是涉及到说和听两个方面。对语音信号的产生与其表征的分析只有与它的感知过程联系起来，才能更好地解决问题。

通常认为，如果我们对人类听觉系统处理信号的过程了解得越好，我们就越能更好地设计一个能够真正理解语音及语意的系统。

人耳是由外耳、中耳与内耳三个部分构成的。

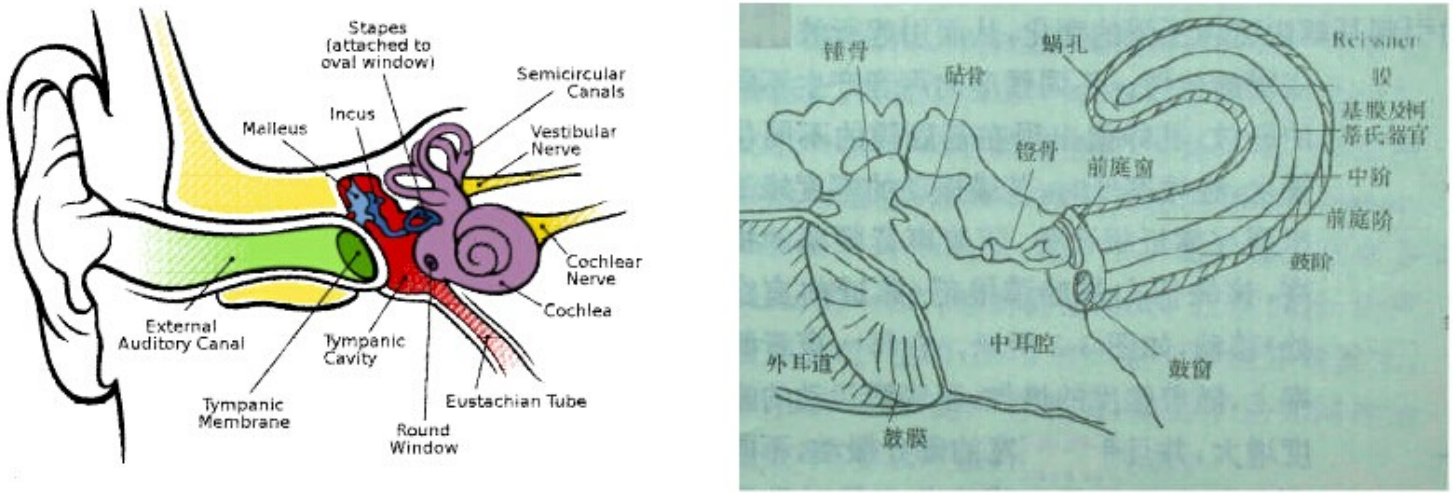


图: 人耳的结构图

外耳 (Outer ear)：

外耳是由耳翼 (耳廓)，外耳道和鼓膜组成。

成年人的外耳道长约 $2.7cm$，直径约 $0.7cm$。外耳道封闭时，最低的共振频率约为 $3060Hz$, 是在语音的频率范围之内。由于外耳道的共振效应，会使声音得到 $10dB$ 的放大。鼓膜呈顶端向内的锥体状，厚度约为 $0.1mm$，面积约为 $69mm^2$，在日常的谈话声作用下，鼓膜位移为 $10^{-8}cm$。

一般认为外耳在对声音感知中的作用为：一是对声源的定位，二是对声音的放大。除了外耳道的共振可导致声音的放大之外，头的衍射效应也会增加鼓膜处的声压，总共可以使声音得到约 $20dB$ 的放大。

中耳：

中耳的总容量约为 $2cm^3$，内含由三块听小骨构成的链，即锤骨，砧骨 (zhen1) 及镫骨。

其中锤骨与鼓膜相接触，镫骨则与内耳的前庭窗 (Oval window) 相接触。中耳的作用：一是进行声阻抗的变换，即将中耳两端的声阻抗匹配起来。另一个是保护内耳。在一定的声强度范围内，听小骨实现声音的线性传递，而在特强声时，听小骨实现非线性传递。

内耳:

内耳 (Inner ear) 的主要构成部分是耳蜗 (Cochlea)。它是听觉的受纳器，把声音通过机械变换产生神经发放信号。

耳蜗长约 $3.5cm$，最宽处约 $0.32cm$，呈螺旋状盘绕 $2.5 \sim 2.75$ 圈。它是一根密闭的管子，内部充满了淋巴液。

耳蜗由三个分隔的部分组成的。

- 鼓阶：靠近耳蜗外轮廓的部分称为鼓阶。鼓阶与中耳的交界处有一个面积为 2mm2 的窗，称为鼓窗
- 前庭阶：靠近耳蜗内轮廓的部分称为前庭阶。鼓阶和前庭阶这两个部分在耳蜗的顶端即蜗孔处是相通的。前庭阶与中耳的镫骨通过前庭窗（卵圆窗）相接触，该窗的面积约为 $3mm^2$。
- 中阶：处在鼓阶和前庭阶之间的部分称为中阶，又称为耳蜗管。中阶的底膜称为基底膜。基底膜上有柯蒂氏器官，它是重要的声音传感装置。

基底膜上的柯蒂氏器官是由耳蜗覆膜、外毛细胞及内毛细胞构成。

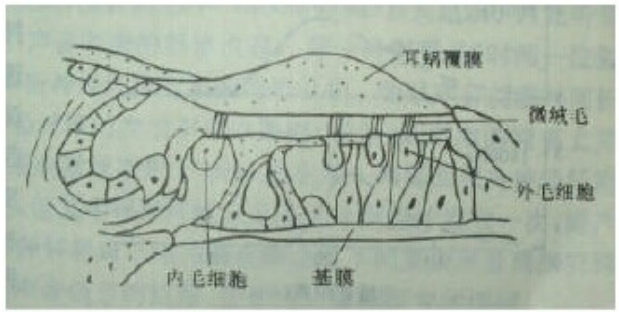


图: 柯蒂氏器官示意图

柯蒂氏器官中毛细胞上的微绒毛受到耳蜗内流体速度变化的影响，从而引起毛细胞膜两边电位的变化，在一定的条件下造成听神经的发放或抑制状态，听神经的这个发放或抑制信号传送到大脑的听觉区域就形成了听觉。

基底膜的结构特点：

1. 尽管耳蜗自身从窗口至蜗顶是越来越小，但是基底膜却越来越宽阔。即靠近前庭窗的部分窄，在蜗顶处宽。
2. 基底膜的韧性越往蜗顶越弱。即靠近前庭窗的部分硬，在蜗顶孔处软。以上基底膜的这两个特点（膜的加宽，质量的增加和韧性的减弱）使得：(1) 蜗顶区对低频声较为敏感。
3. 正圆窗和卵圆窗处的基底膜对高频声较敏感。

基底膜的运动特点：

1. 基底膜上的传输的是行波。只需几毫秒便从基底膜的一端传向顶端。
2. 行波的振幅正与比声强成正比。
3. 由于基底膜的弹性不同，使得低频音的振幅最大位置在蜗顶处，而高频音的振幅的最大位置在靠近卵圆窗处。

基底膜上的毛细胞分布

基底膜的毛细胞分为外侧毛细胞和内侧毛细胞两种。其中内侧毛细胞有 $3 \sim 5$ 排，约 20000 个。内侧毛细胞是单排的，约有 3500 个。毛细胞在根部分布的密，在蜗顶毛细胞的数量较少。

连接内侧和外侧毛细胞到脑的神经纤维约有 30,000 个左右。基底膜的任何运动都直接通过各种各样的应力和张力作用于毛细胞上，激起神经活动，发生电冲动，传向听神经。

总结:

在基底膜不同位置的毛细胞具有不同的电学与力学特性。在耳蜗的根部，基底膜窄而劲度强，外毛细胞及其绒毛短而有劲度；在蜗顶处，基底膜宽而柔和，毛细胞及其绒毛也较长而柔和。由于这种结构上的差别，因而它们具有不同的机械谐振特性和电谐振特性。据认为，这种差别在确定频率选择特性时可能是最重要的因素。

每条传入听神经纤维因为只连接到一个基底膜的内毛细胞上，因此它只对一定频率范围内的声音信号发生响应，即具有调谐性质；并且对某一特定频率最为敏感，该频率称之为特征频率（Characteristic frequency）。

内耳听神经的发放有如下特点：

- 在无声音刺激时，具有自发活动，其速率从每秒 0.5 次到 100 次不等且具有随机性。在强刺激建立时发放率可达每秒 1000 次，而在稳定强刺激时发放率逐渐下降到每秒 150 次左右。
- 每次发放后，约有 1ms 的不应期，然后回到正常的速率。
- 线性半波整流特性，即发放概率几乎是刺激波形“正”部分的复本，而“负”部分不发放。
- 随刺激声音强度的增加，发放率随之增加，当声音强度达到一定水平时，发放率达到饱和。

1. $Phon$（方）：这是一种用来表示纯音响度的单位。其参考声压级为 $1dB_{spl}$ 的 1000Hz 单频纯音的响度为 1 个 $Phon$（0dB）(P)。
2. $Song$（宋）：定义声压级为 $40dB_{spl}$ 的 1000Hz 单频纯音的响度为 1 个 $Song$(S)。

注意 $Phon$ 实际上是以分贝为单位。$Song$ 和 $Phon$ 的关系不是线性关系。　　当 $Phon > 40dB$ 时

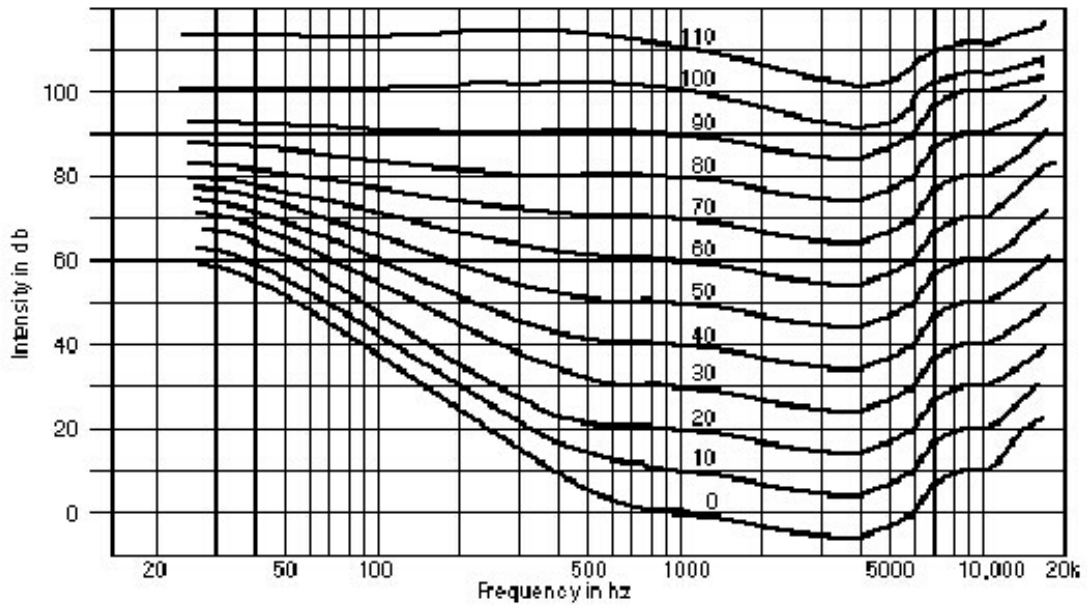
$$S = 2^{(P-40)/10} \tag{11}$$

或者为：

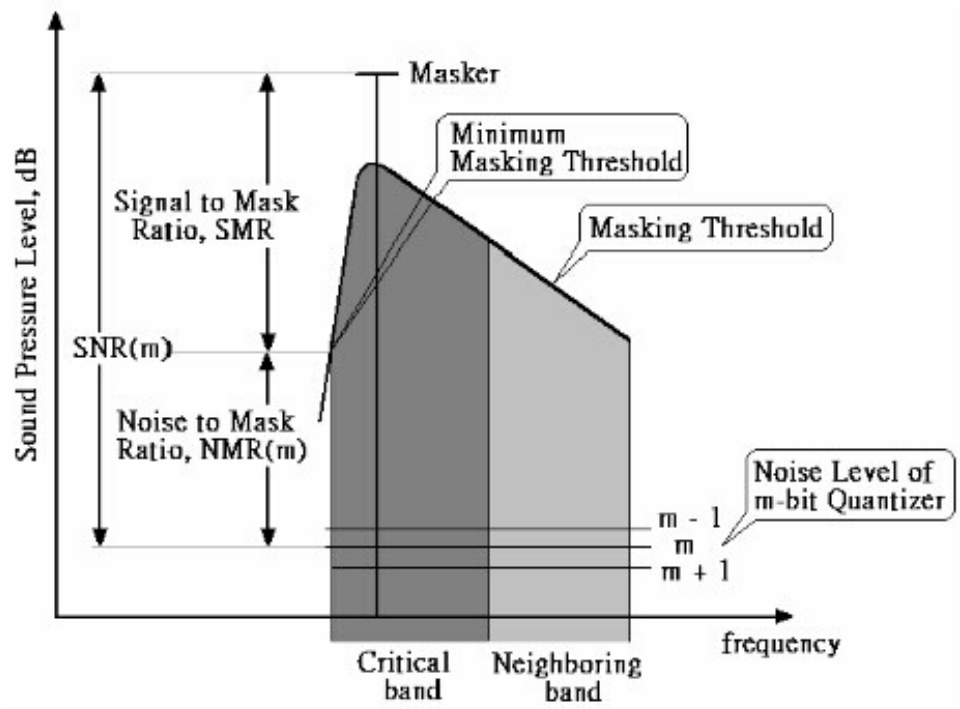
$$P = 33.33log(S) + 40 \tag{12}$$

从上式不难看出，1 Sone 相当于 40 Phone，响度 S 增加一倍相当于响度级 P 增加 10Phon。

声音的响度是与人耳主观感觉有关的单位。其参照声为 1000Hz 的声压级。



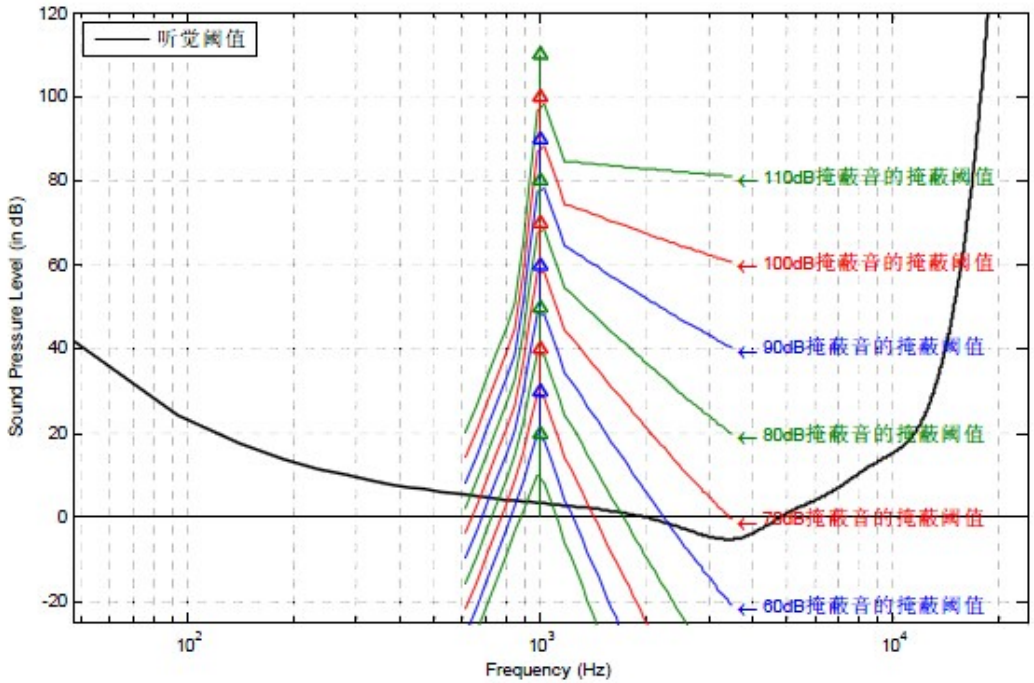
一个较弱的声音 (被掩蔽音) 的听觉感受被另一个较强的声音 (掩蔽音) 影响的现象称为人耳的“掩蔽效应”。注意：声强级大的音称为掩蔽音。



在掩蔽情况下，提高被掩蔽弱音的强度，使人耳能够听见时的闻阈称为掩蔽门限。

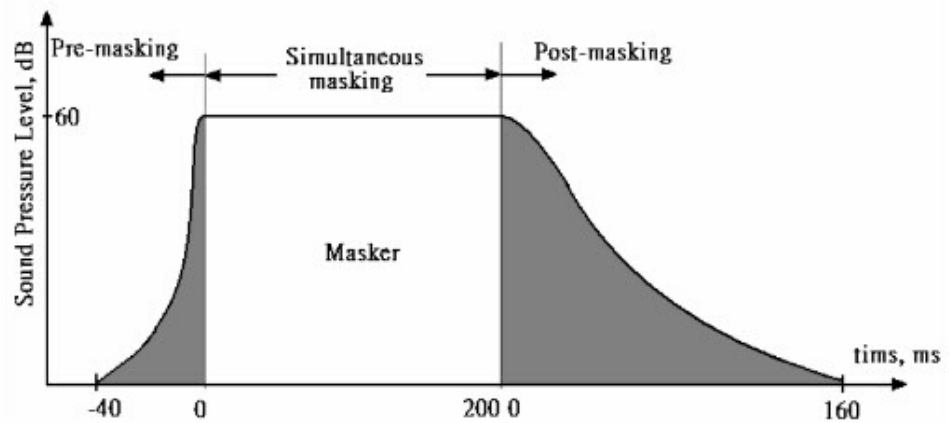
掩蔽效应可以分为在频域上的掩蔽 (Simultaneous Masking) 和时域上的掩蔽 (Temporal Masking) 两种情况。

- 频域掩蔽曲线



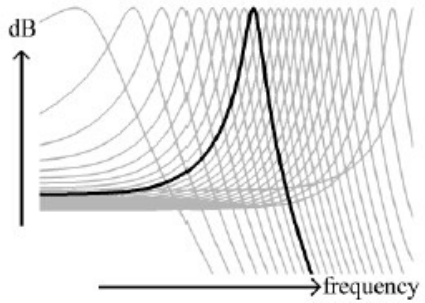
特点：不对称。掩蔽音对低频部分掩蔽作用小，而对高频音掩蔽作用大。

- 时间掩蔽：前向掩蔽 (Pre-Masking) 和后向 (Post-Masking) 掩蔽



前向掩蔽：当前强掩蔽音对其前数毫秒内的声音的掩蔽现象。3ms—20ms
后向掩蔽：当前强掩蔽音对其后数百毫秒内的声音的掩蔽现象。50ms—100ms

- 噪声的存在也会影响到纯音的接收，即对纯音产生掩蔽。
- 为了描写这种掩蔽的效果，引入了临界带宽 (Critical band) 的概念。
- 一个纯音，可以被以它为中心频率，并且具有一定频带宽度的连续噪声所掩蔽，即称这一带宽为临界带宽。



临界带宽编号 Z（Bark）与频率 $f(Hz)$ 之间的关系近似为

$$Z \cong 26.81 f/(1960+f) - 0.53 \tag{13}$$

表 1 人耳听觉的临界带

临界带编号	低频频率 (Hz)	中心频率 (Hz)	高频频率 (Hz)	带宽 (Hz)	临界带编号	低频频率 (Hz)	中心频率 (Hz)	高频频率 (Hz)	带宽 (Hz)
1	0	50	100	100	2	100	150	200	100
3	200	250	300	100	4	300	350	400	100
5	400	450	510	110	6	510	570	630	120
7	630	700	770	140	8	770	840	920	150
9	920	1000	1080	160	10	1080	1170	1270	190
11	1270	1370	1480	210	12	1480	1600	1720	240
13	1720	1850	2000	280	14	2000	2150	2320	320
15	2320	2500	2700	380	16	2700	2900	3150	450
17	3150	3400	3700	550	18	3700	4000	4400	700
19	4400	4800	5300	900	20	5300	5800	6400	1100
21	6400	7000	7700	1300	22	7700	8500	9500	1800
23	9500	10500	12000	2500	24	12000	13500	15500	3500
25	15500	19500	-	-					

音阶：

把 C、D、E、F、G、A、B 等各音中的某一个音作为中心，由它开始由低至高（或由高到低）按顺序排列起来，这个音的序列由于像梯子一样，逐级向上或向下，所以叫音阶。

自然音阶：

没有升（#）降（b）的音阶。

音名	C	D	E	F	G	A	B
唱名	Do	Re	Mi	Fa	Sol	La	Ti
简谱	1	2	3	4	5	6	7

- 3-4 与 7-1 之间音程为半音，其于为全音（两个半音）。
- 半音在指板上相距一琴格。

- 全音在指板上相距两琴格。

拨一根弦与拨一根长度只有一半的弦（即长度比为 1/2）其振动发出同样的音调，并构成了一个音阶的长度。较短的弦比原来的弦每秒振动的次数多一倍。

【注】术语第八音（octave）来自拉丁词 8。一个全音阶有 7 个不同的音调，从 C 调到 B 调和高 C 调，第八音则是第八音调）。

变化音阶：

全部由半音（即有升降调）组合而成，又称半音音阶。

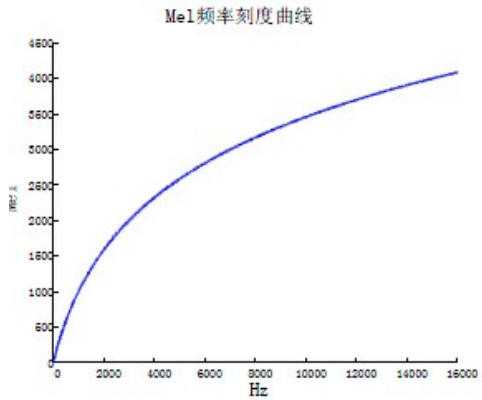
音名	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
唱名	Do	Di	Re	Ri	Mi	Fa	Fi	Sol	Si	La	Li	Ti
简谱	1	#1	2	#2	3	4	#4	5	#5	6	#6	7

【注】（#1 = b2 . #5 = b6 . #4 = b5 . #2 = b3 . #6 = b7 ）称为同音异名。A4(键盘中央 C 上面的 A)440Hz。8 度音频率差 2 倍，8 度音程包括 12 个音（7 个全音，5 个半音），它们在听觉上是等间距的，前后音符的频率关系为 $2^{1/12}$ 倍。

美阶

音调是听觉分辨声音高低时，用于描述这种感受的一种特性。对于频率低的声音，听起来感觉就“低”，而频率高的声音，听起来感觉就“高”。但是音调与声音的频率并不是成正比关系。为了描写音调，采用了美（Mel）标度，这就是所谓的美阶。

定义：一个高于听觉阈 40dB、频率为 1KHz 的纯音所产生的音调定义为 1000Mel。这样如果一个纯音听起来比 1000Mel 的声音调子高一倍，则其音调为 2000Mel。



$$T_{Mel} \cong 3322.23 log_{10}(1 + 0.001 f_{Hz}) \tag{14}$$

德国物理学家 Helmholtz 提出音调的高低是由在基底膜的最大振动位置来决定。变化 1mel 约等于该振动在基底膜上穿越 12 个神经元。

耳朵对频率变化的敏感性度量单位：最小可觉差（jnd）。

100Hz 时变化约 3%，1000Hz 时变化约 0.3%，人耳就可以察觉出变化。

1 mel 约为 12 个神经元，0.23 最小可觉差，0.009 临界带。

1 最小可觉差约为 52 个神经元，4.3mel，0.04 临界带。

1 临界频带约为 1300 个神经单元，108mel，2.5 最小可觉差。