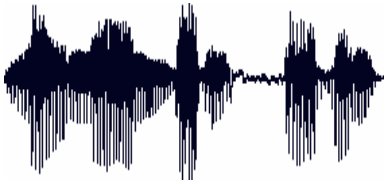
第二章　语音信号处理的基础

为什么需要语音信号处理？



语音信号

- 内容（Content） → 语音识别
- 说话人（Speaker） → 声纹识别
- 性别（Gender）
- 情感（Emotion） → 情感识别
- 噪音（Noise） → 语音增强
- 韵律（Prosody）
- 语言（Language） → 语言识别
- 口音（Accent） → 口音识别
- 空间信息（Spatial information）
- …

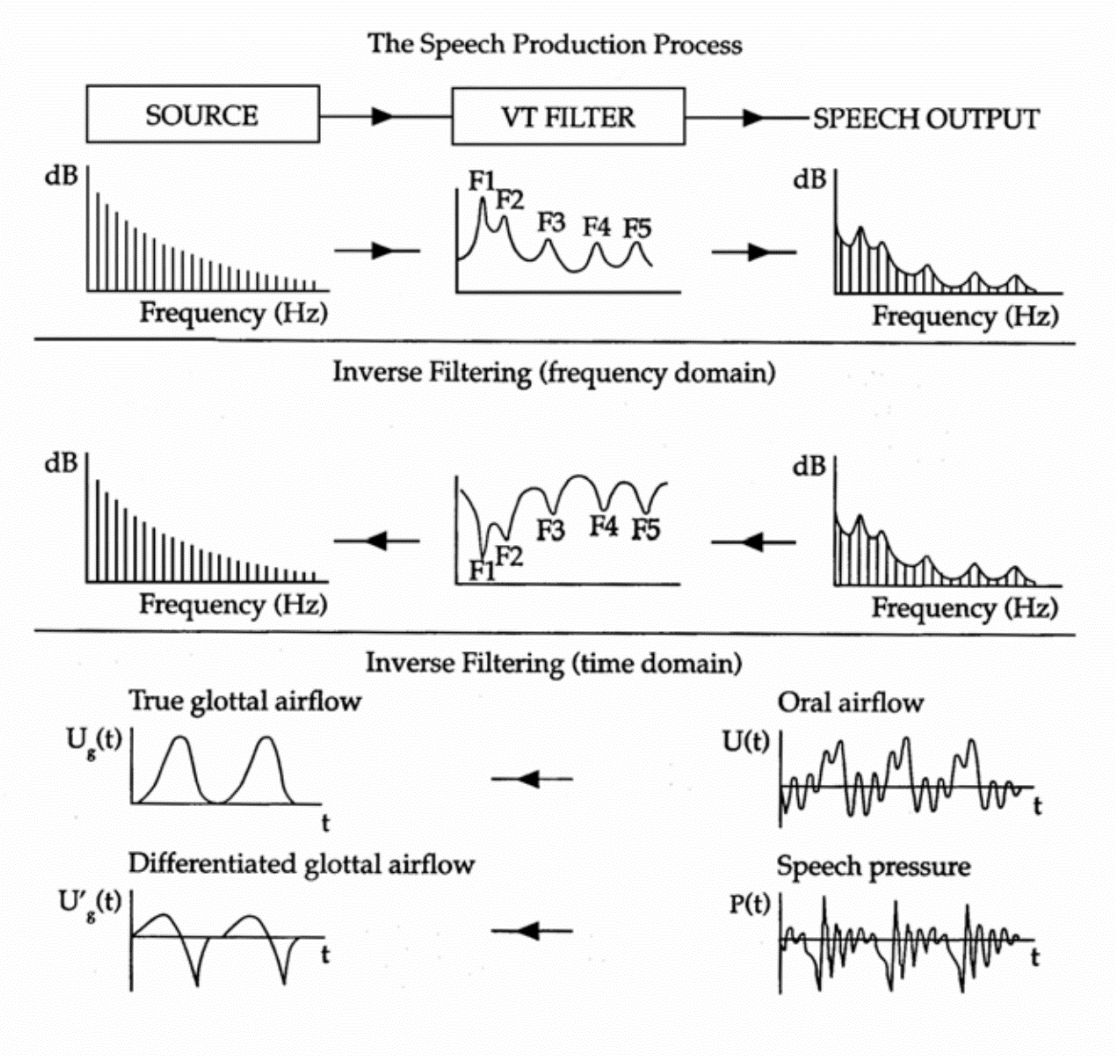
语音是经过语言调制的声学信号

- 语音信号处理主要关注语音信号的调制和解调的过程，但较少关注语言规划和语义理解。
- 语音的调制过程是由发音器官运动生成的动态声道对浊音声源或清音声源进行滤波实现的，即声源-滤波器模型；
- 语音的解调过程是由内耳的基底膜将声波振动分解为不同频率的神经电信号传递给大脑听觉皮层进行感知的过程。
- 在长期的语音信号处理方法的研究过程中，人们主要遵循语音产生机理解声道特性和声源特性，以及基于人的语音感知机理去分析语音特征。本章将沿着这个思路介绍语音信号处理的基础知识和方法。

语音信号处理的目的

- 输入和脉冲响应已知，求输出
- 输入和输出已知，求系统特性
- 输出已知，求系统特性和输入

-- 语音信号处理多属于第3种情况



第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
  - 2.1.1 语音信号的数字化（**了解**）
  - 2.1.2 语音信号的时域分析（**基本概念**）
  - 2.1.3 语音信号的频域分析（**基本概念**）
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
- 2.4 基于语音感知机理的特征分析方法

第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
  - 2.1.1 语音信号的数字化
  - 2.1.2 语音信号的时域分析
  - 2.1.3 语音信号的频域分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
- 2.4 基于语音感知机理的特征分析方法

信号为什么需要离散化？

- 因为计算机不可能也没必要用无穷大的内存存储连续模拟信号，因此需要将连续模拟信号在时间和幅度上进行离散化。

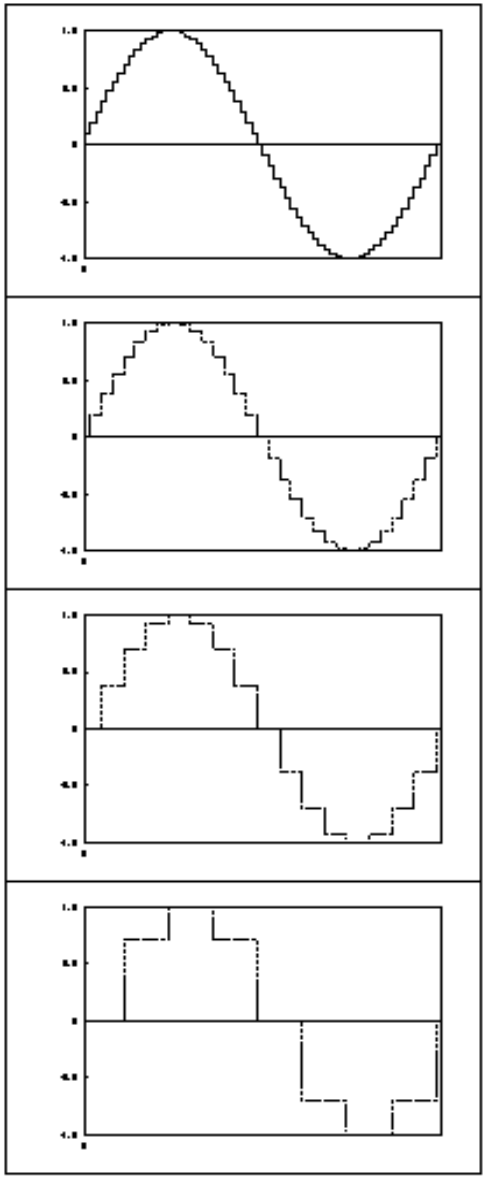
- 时间上的离散化称为采样（或抽样）

  $s(n) = s_a(nT)$，　$-\infty < n < \infty$

  T：采样周期；$f = \frac{1}{T}$：采样频率

- 高采样频率可以将信号波形表达地更加准确

- 增加量化的bit数可以减少振幅的量化误差



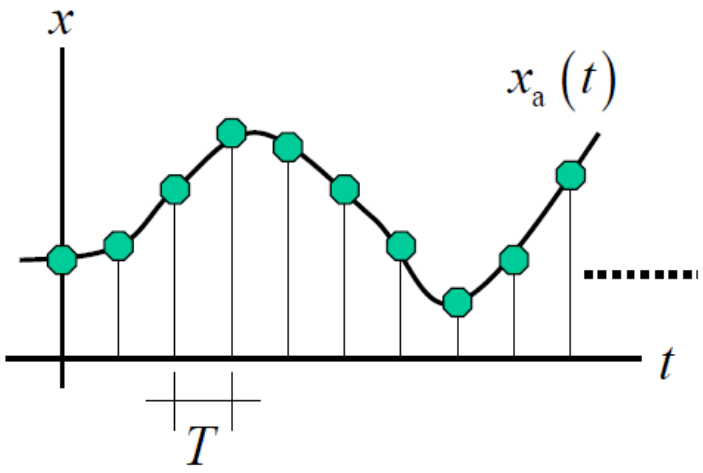
64 samples/cycle

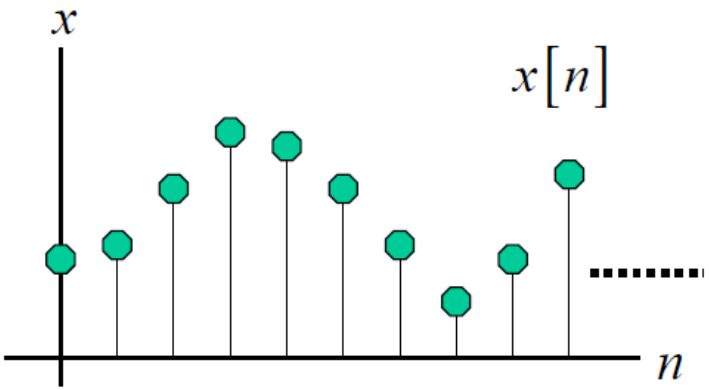
32 samples/cycle

16 samples/cycle

8 samples/cycle

语音信号的采样





采样是以一定的间隔获得一个函数的值的操作。

$$x_s[n] = x_a(nT), \quad -\infty < n < \infty$$

$$= \int_{t=-\infty}^{\infty} x_a(t)\delta(t - nT)dt$$

$x_a(t)$: 模拟信号
(连续的时间信号)

T: 采样周期

δ: 狄拉克delta函数

奈奎斯特频率（基本概念）

- 奈奎斯特频率：离散信号系统采样频率的一半。

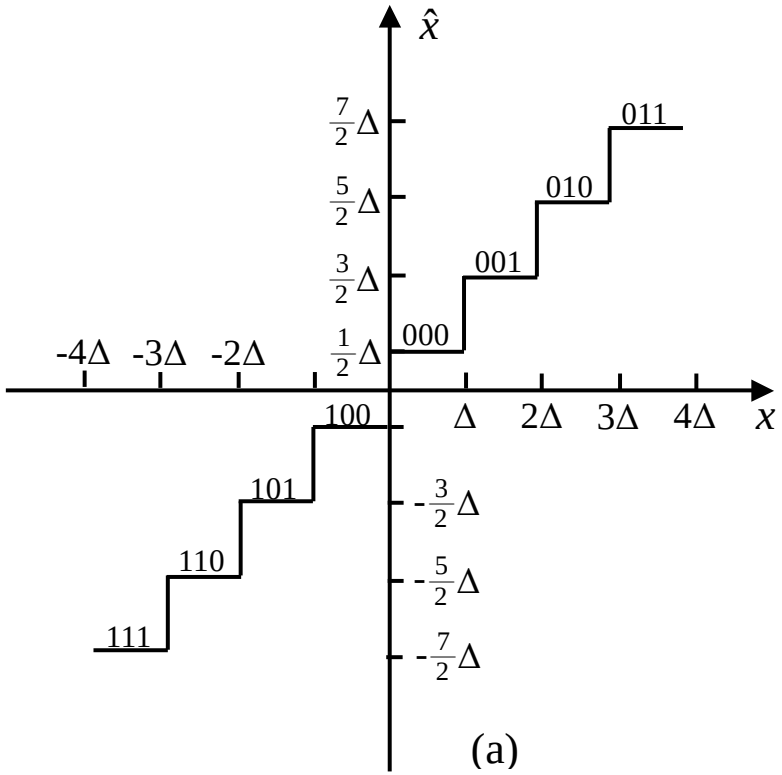
- 奈奎斯特－香农采样定理：对于一个给定的采样频率f_s，完全重构的频带限制为$B \leq f_s/2$。也就是所，只要离散系统的奈奎斯特频率高于被采样信号的最高频率或带宽，就可以避免混叠现象。

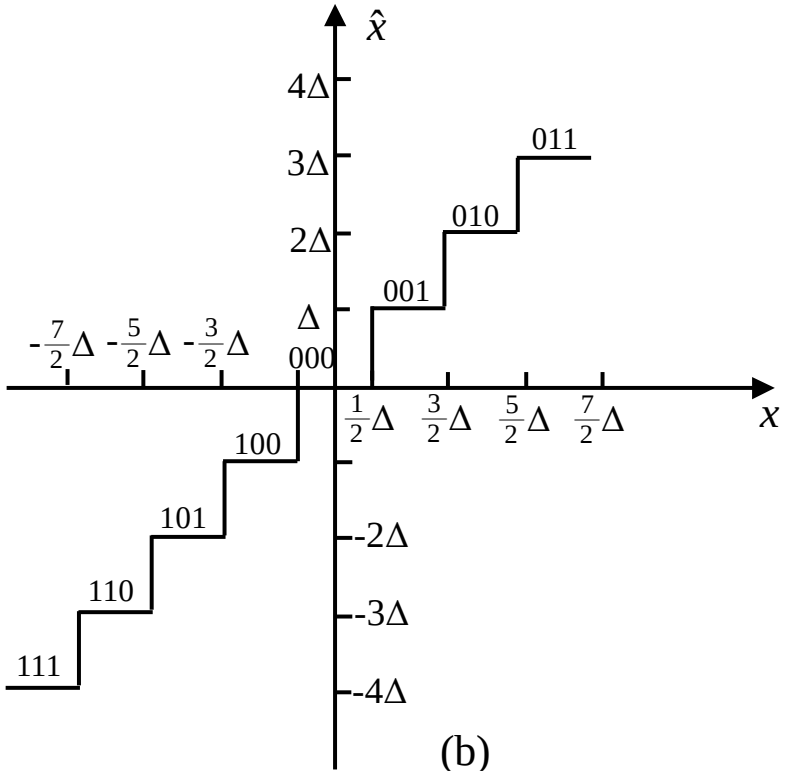
- 奈奎斯特－香农采样定理的名字是为了纪念哈里·奈奎斯特和克劳德·香侬。

常用的采样频率：8k Hz，16k Hz，44.1k Hz

人的听觉范围：20-20000 Hz

语音信号的量化





量化误差： $e(\mathrm{n}) = \hat{x}(n) - x(n)$

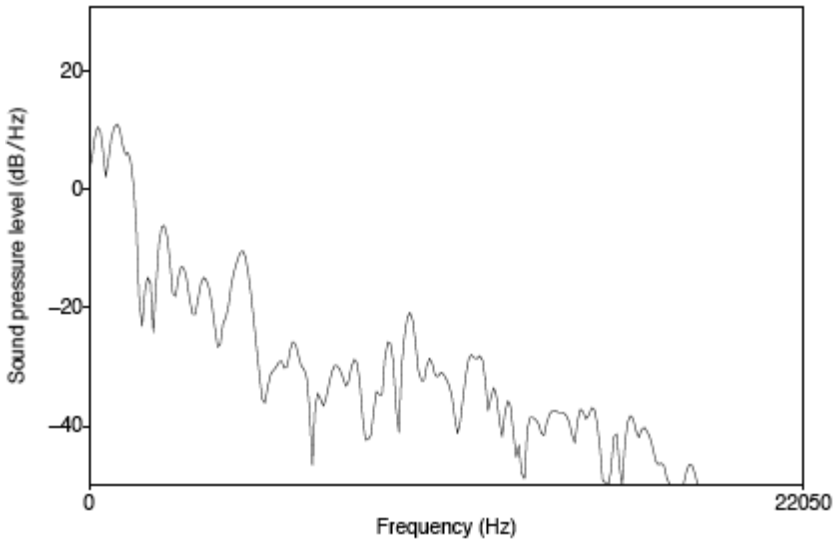
常用的量化比特：16比特

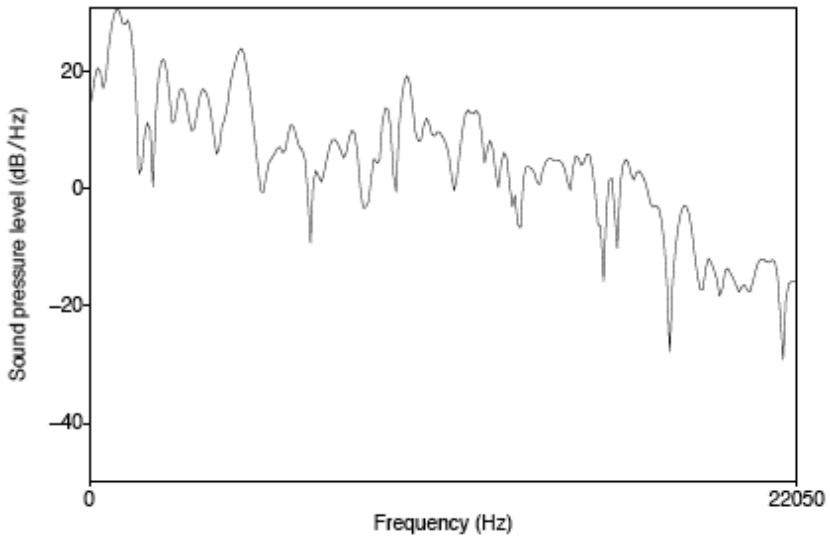
语音信号预处理-预加重(Pre-emphasis)

- 预加重: 提升高频部分，使信号的频谱变得平坦
- Q:理由?
- A: 语音信号的平均功率谱受声门激励和口鼻辐射的影响，高频端大约在800 Hz以上按照-6 dB/倍频程跌落

语音信号预处理-预加重(续)

- 预加重前后频谱比较（元音［aa］）

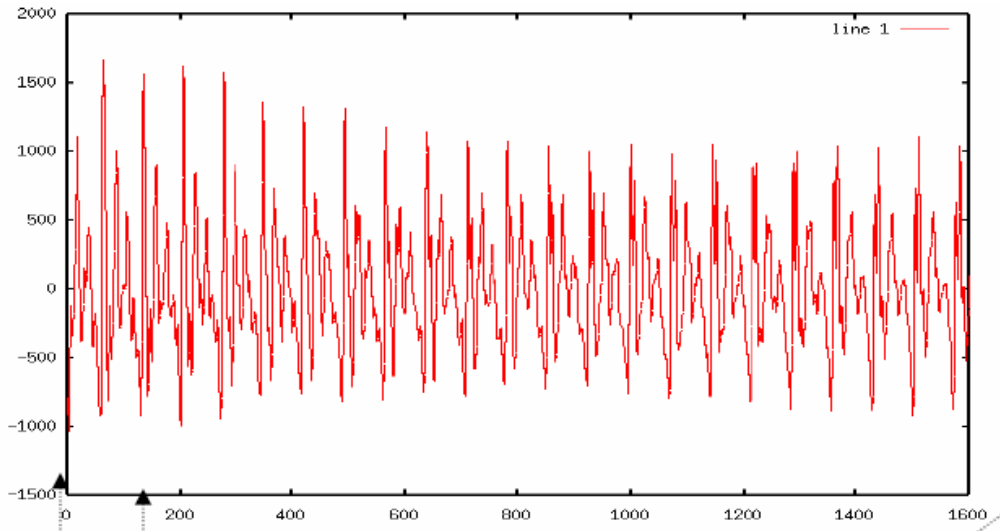




$H(z) = 1 - az^{-1}$

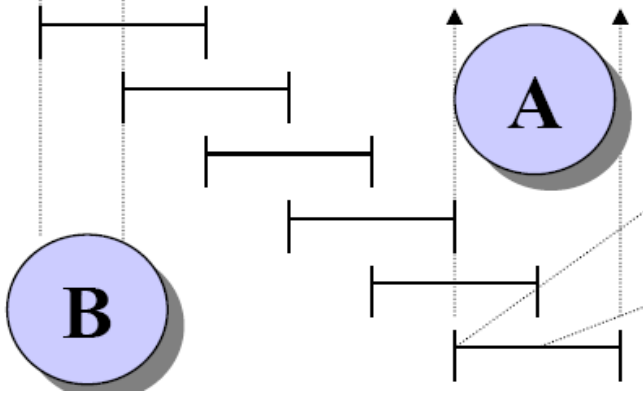
a的经典取值为0.94~0.99

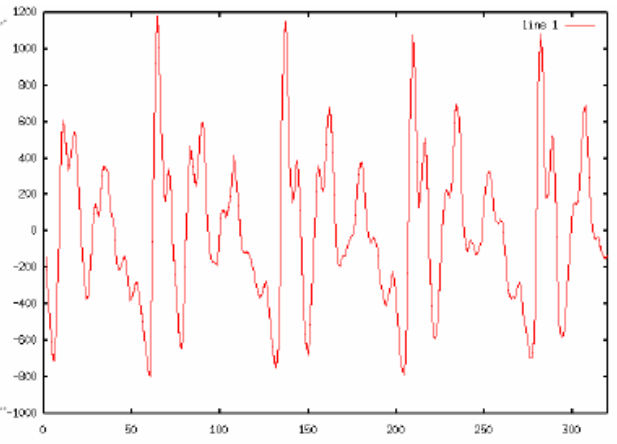
语音信号预处理-分帧(Framing)（基本概念）



A ~ 20 – 25 ms 帧长

B ~ 10 ms 帧移





非周期的语音信号 ⟹ 准周期信号

语音信号预处理-加窗(Windowing)（基本概念）

目的：为了使时域信号更好地满足离散傅里叶变换的周期性要求，减少泄漏。

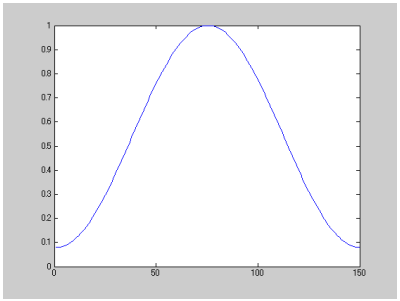
加窗后的语音信号：

$$x'(n) = x(n)w(n)$$

语音信号预处理-加窗(续)（基本概念）

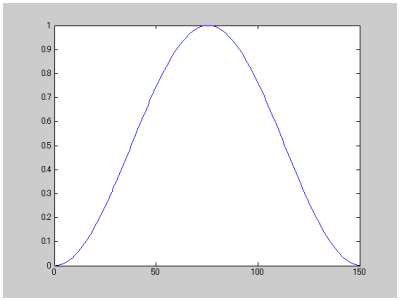
- 汉明窗（Hamming window）：

$$w(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)] & 0\le n\le N-1\\ 0 & \text{其它}\end{cases}$$



- 汉宁窗（Hanning window）：

$$w(n)=\begin{cases}0.5[1-\cos(2\pi n/(N-1))] & 0\le n\le N-1\\ 0 & \text{其它}\end{cases}$$



- 矩形窗（Rectangular window）：

$$w(n)=\begin{cases}1 & 0\le n\le N-1\\ 0 & \text{其它}\end{cases}$$

语音信号预处理-分帧(Framing)

为什么分帧的窗长一般取20-25 ms？

提示：基音范围

80~500 HZ，儿童和青年女性偏高，成年男性偏低

对应周期2ms-12.5 ms

窗长：1-7个基音周期

第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
  - 2.1.1 语音信号的数字化
  - 2.1.2 语音信号的时域分析
  - 2.1.3 语音信号的频域分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
- 2.4 基于语音感知机理的特征分析方法

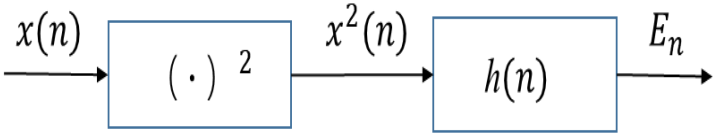
时域信号处理方法-短时能量

- 短时能量（short-time energy）

E_n: 信号的第n个点开始加窗函数的短时能量

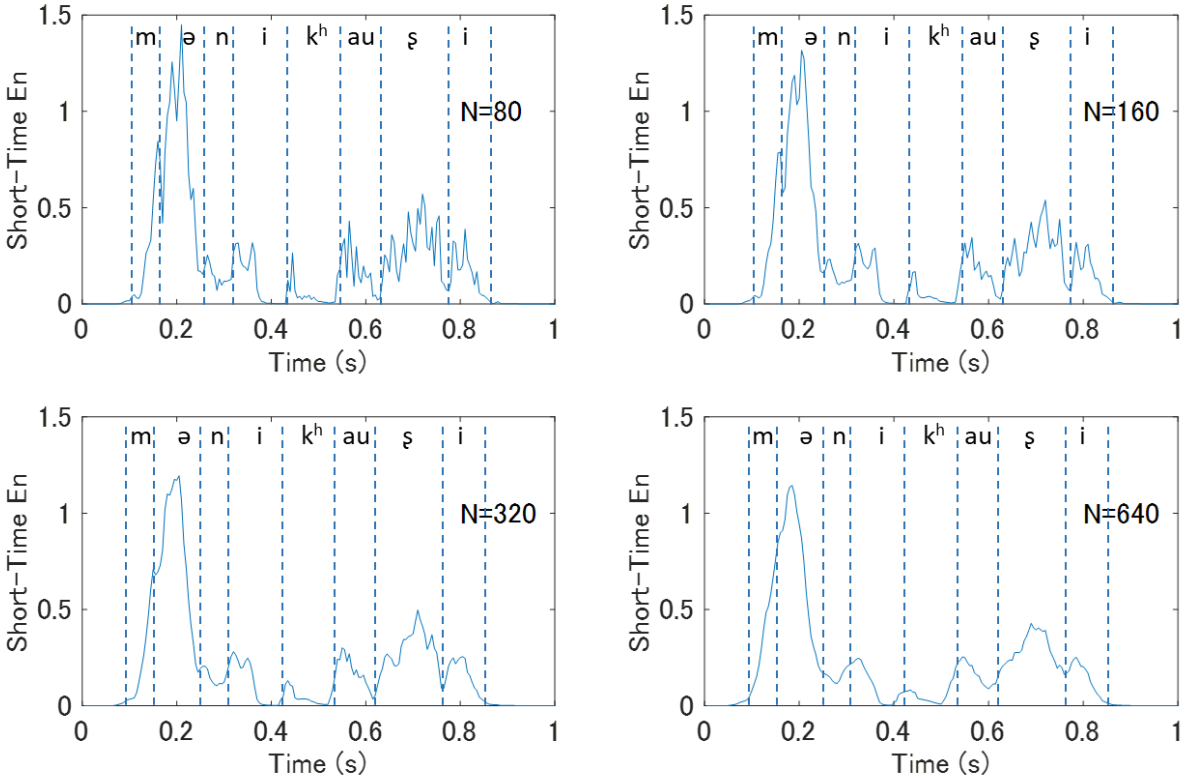
$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m)$$

$h(n) = w^2(n)$　　其中，$w(n)$ 是窗函数



短时平均能量的计算

时域信号处理方法–短时能量（续）



不同宽度矩形窗的短时能量函数

应用：浊音和清音的区分、语音端点检测（有声段和无声段的判定）、声母与韵母的分界、特征分析（语音情感分类的一种特征）等

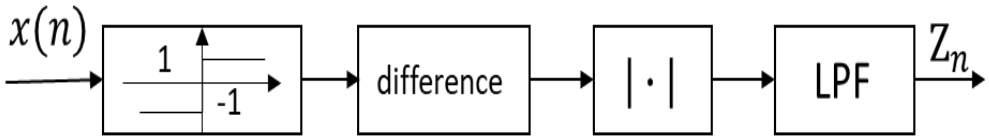
时域信号处理方法-短时间平均过零率

- 短时间平均过零率 （zero-cross rate：ZCR）：
  信号采样点符号变化的次数

其频率表征可由下式计算

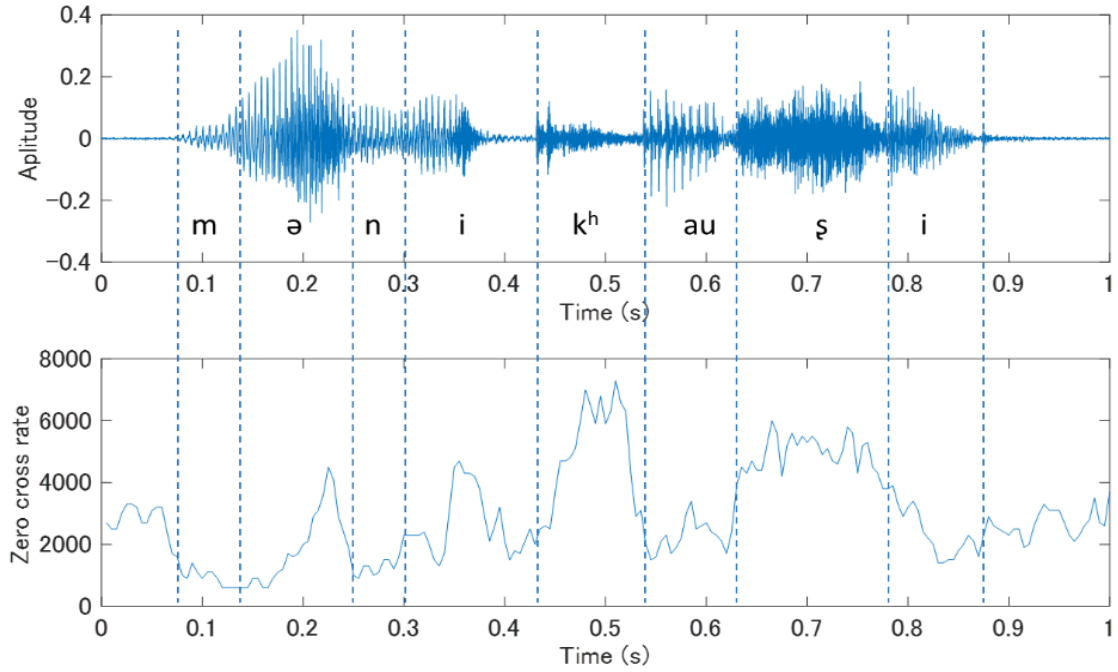
$$Z = \frac{f_s}{N}\sum_{n=0}^{N-1}|sign[x(n)] - sign[x(n+1)]|$$

其中，f_s为采样频率，N为窗的点数，$sign[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases}$



短时平均过零率的计算

时域信号处理方法-短时间平均过零率（续）



ZCR反映了包括主能量在内的区域的平均频率：

- 浊音：大约 1400 Hz
- 清音：大约 4500 Hz

应用：浊音和清音的区分、语音端点检测（短时平均能量和ZCR的结合）等

时域信号处理方法-短时自相关函数

- 自相关函数（Autocorrelation）：
  自相关函数是测定相同信号在时域内的相似度

$$R(i) = \sum_{n=-\infty}^{\infty} x(n)x(n-i), \qquad i = \{0,1,2,\dots\}$$

- 短时自相关函数（Short-time autocorrelation）：
  在自相关函数的基础上加短时分析

$$R(m) = \frac{1}{2N+1} \sum_{n=-N+1}^{N-1} x(n)x(n+|m|), (|m| = 0,1,\cdots,N-1)$$

主要用于研究信号本身的同步性，周期性

时域信号处理方法-短时自相关函数（续）

- 性质

  - 对称性

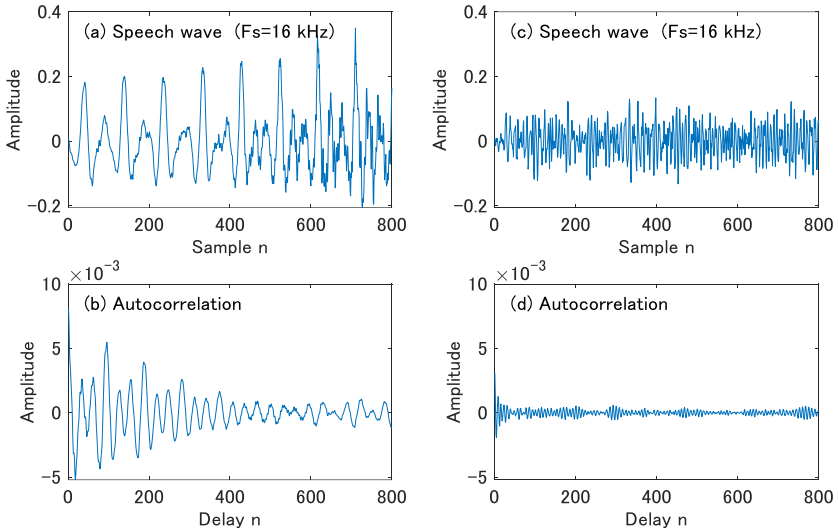
    $$R(m)=R(-m)$$

  - m =0时达最大值

    $$|R(m)| \leq R(0)$$

  - 白噪音的自相关

    连续时间白噪声信号的自相关在$m = 0$处有一个强峰值，而在所有其他$m \neq 0$处值为0。



应用：浊音和清音的区分、浊音周期估计等

短时自相关函数为什么可以应用到以上任务？

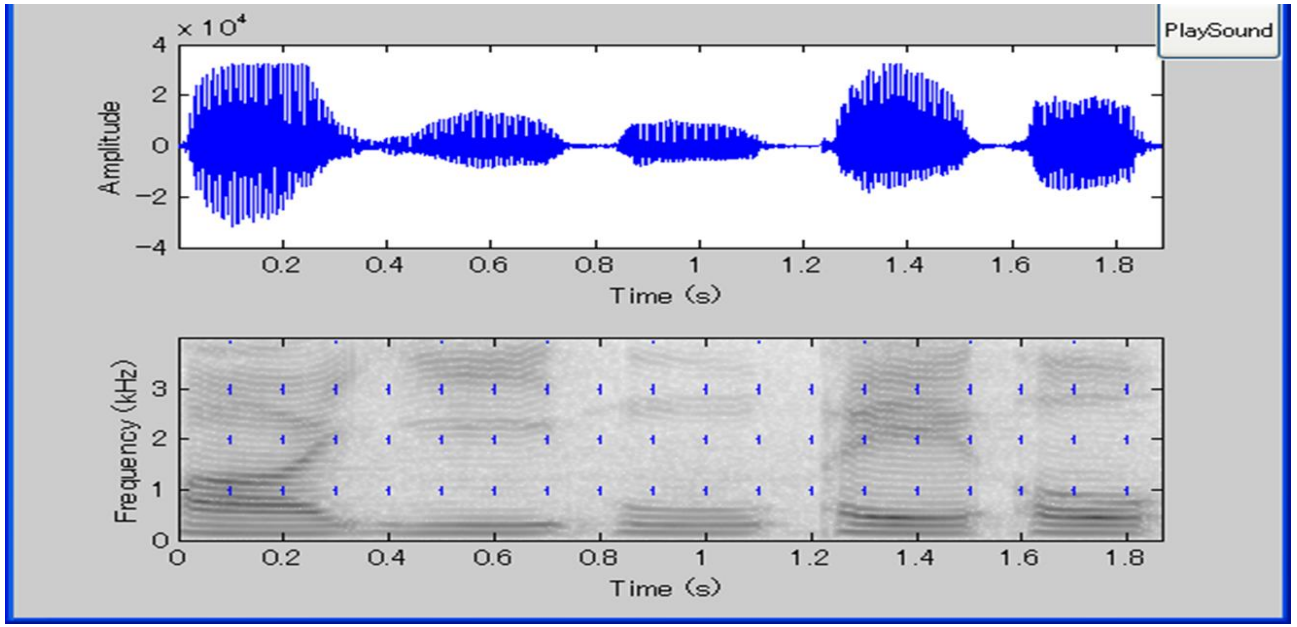
如果$x(n)$是一个周期信号，$R(i)$就会在i等于周期整数倍的点取极大值。

第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
  - 2.1.1 语音信号的数字化
  - 2.1.2 语音信号的时域分析
  - 2.1.3 语音信号的频域分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
- 2.4 基于语音感知机理的特征分析方法

为什么需要频域分析？

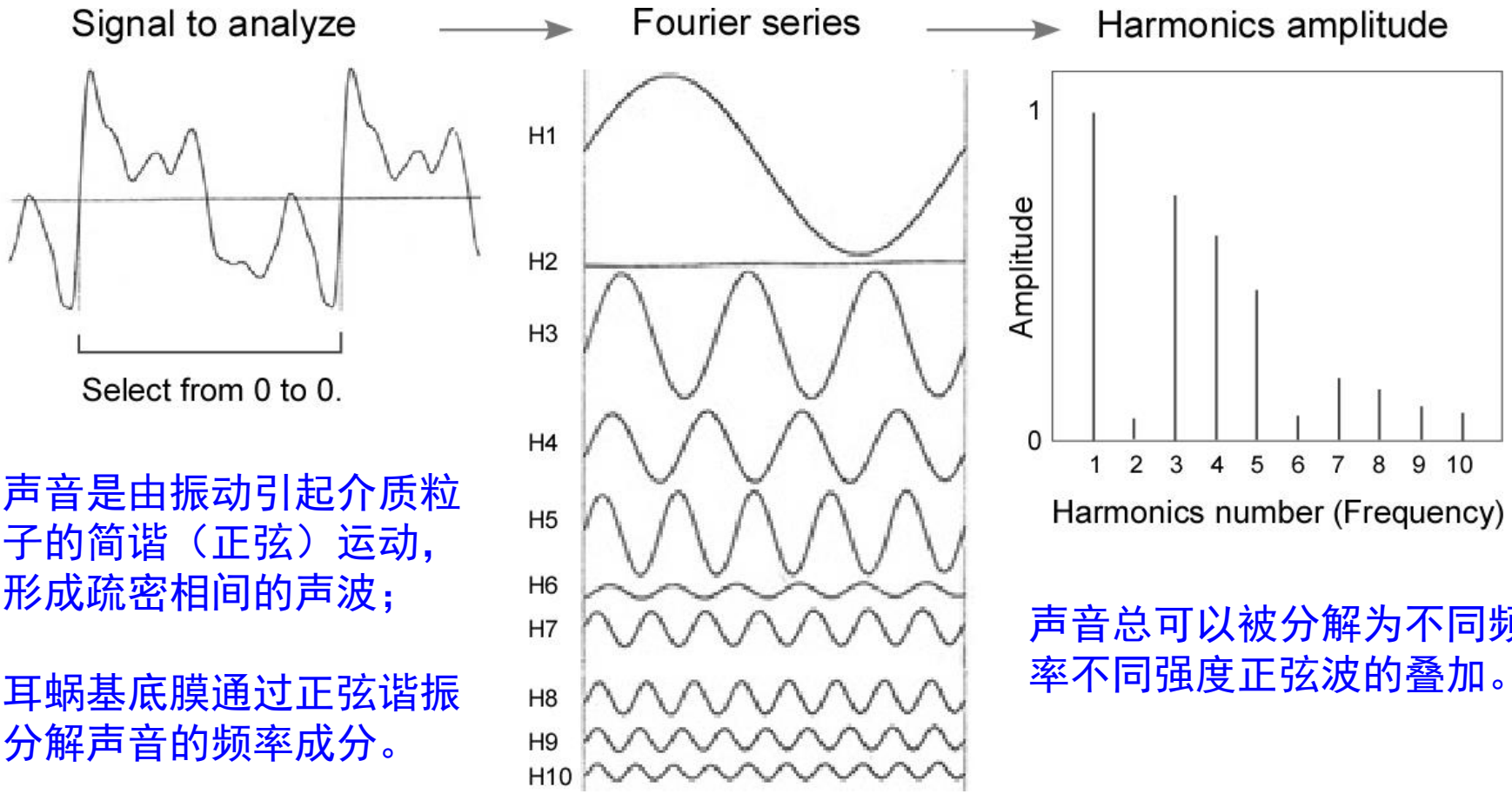
- 人类听觉系统具有频谱分析功能
- 频域的计算更简便：时域的卷积相当于频谱的乘积



语音信号的频谱分析，是认识和处理语音信号的重要方法

谐波分析（了解）

声音取决于谐波模式，早期的时频分析采取谐波分量分解和振幅计算。

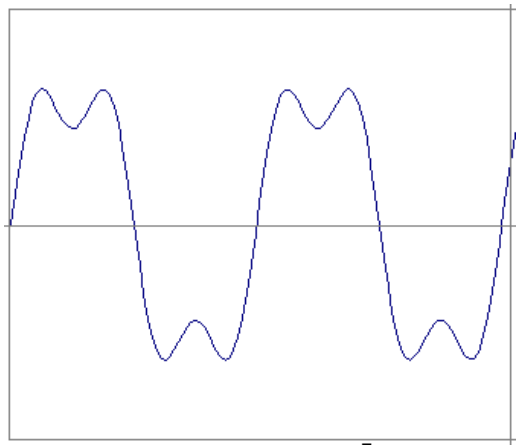


- 声音是由振动引起介质粒子的简谐（正弦）运动，形成疏密相间的声波；
- 耳蜗基底膜通过正弦谐振分解声音的频率成分。

声音总可以被分解为不同频率不同强度正弦波的叠加。

信号的时间和频率特性（了解）

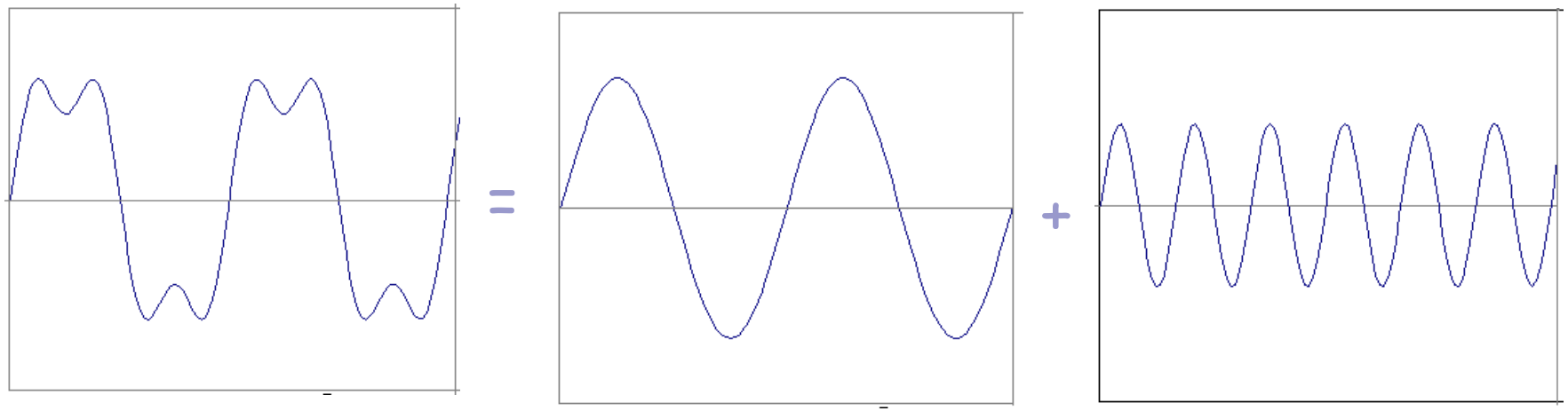
- 例: $g(t) = \sin(2\pi f_0 t) + (1/3)\sin(2\pi (3f_0) t)$



时间和频率特性反映组成该信号的不同频率成分的时间变化

信号的时间和频率特性（了解）

- 例: $g(t) = \sin(2\pi f_0 t) + (1/3)\sin(2\pi (3f_0) t)$



时间和频率特性反映组成该信号的不同频率成分的时间变化

离散傅里叶变换

离散傅里叶变换（Discrete Fourier Transform）

$$X(k)=\sum_{n=0}^{N-1}x(n)\exp\left(-j\frac{2\pi kn}{N}\right):(0\le k\le N-1)$$

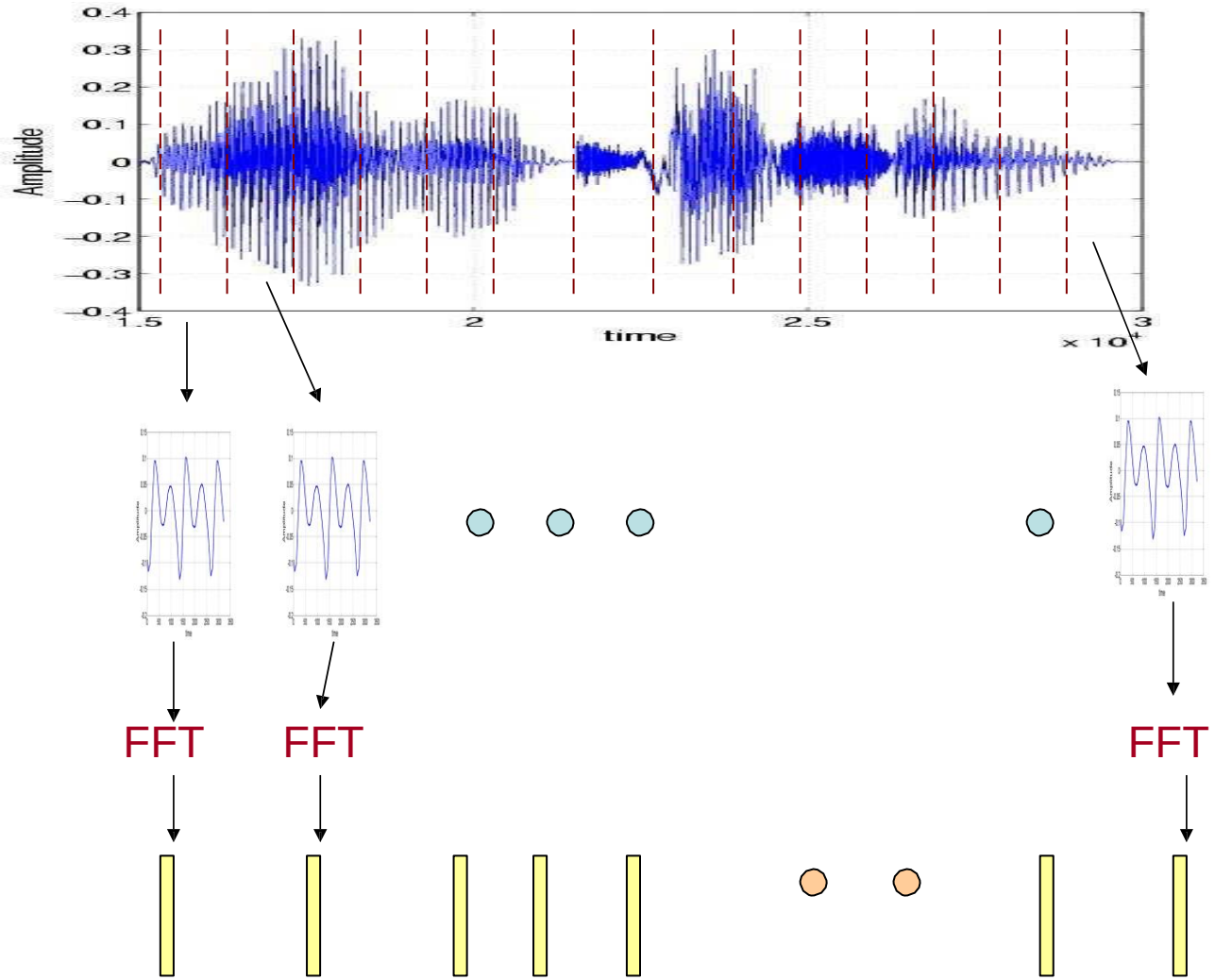
$x(n)$：时间窗内采样的离散语音信号

离散傅里叶逆变换（Inverse Discrete Fourier Transform）

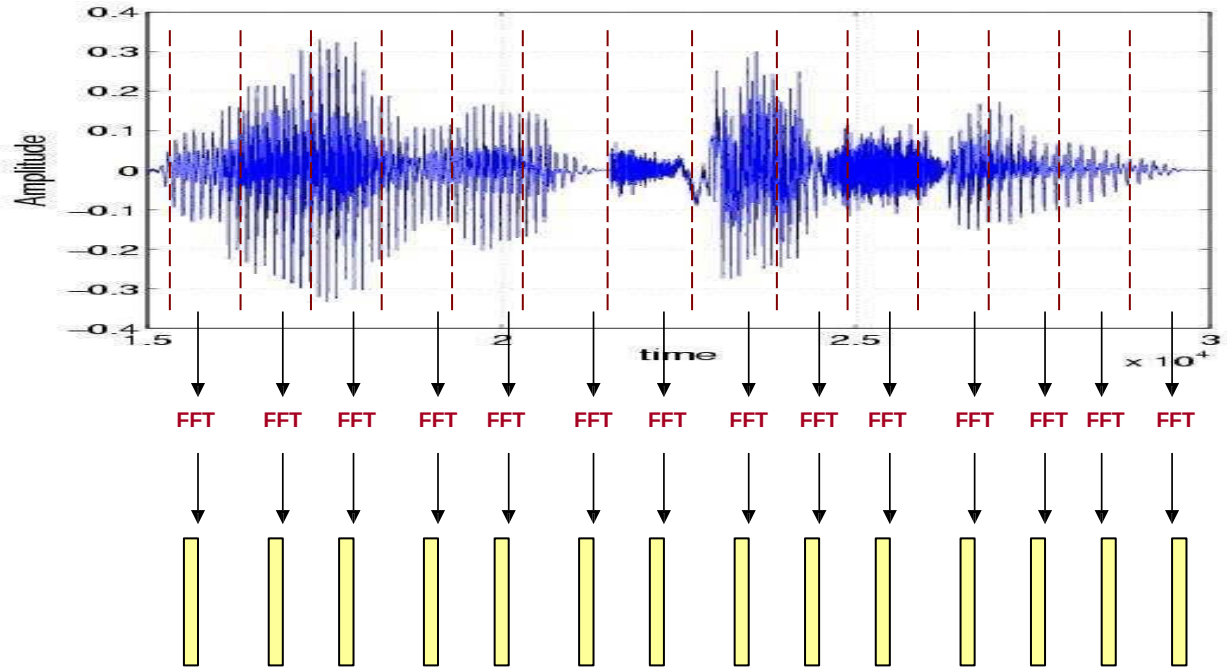
$$x(n)=\frac{1}{N}\sum_{n=0}^{N-1}X(k)\exp\left(j\frac{2\pi kn}{N}\right)\ :(0\le n\le N-1)$$

X(k)：语音信号的离散频谱（复数序列）

从语音信号到频谱向量的序列

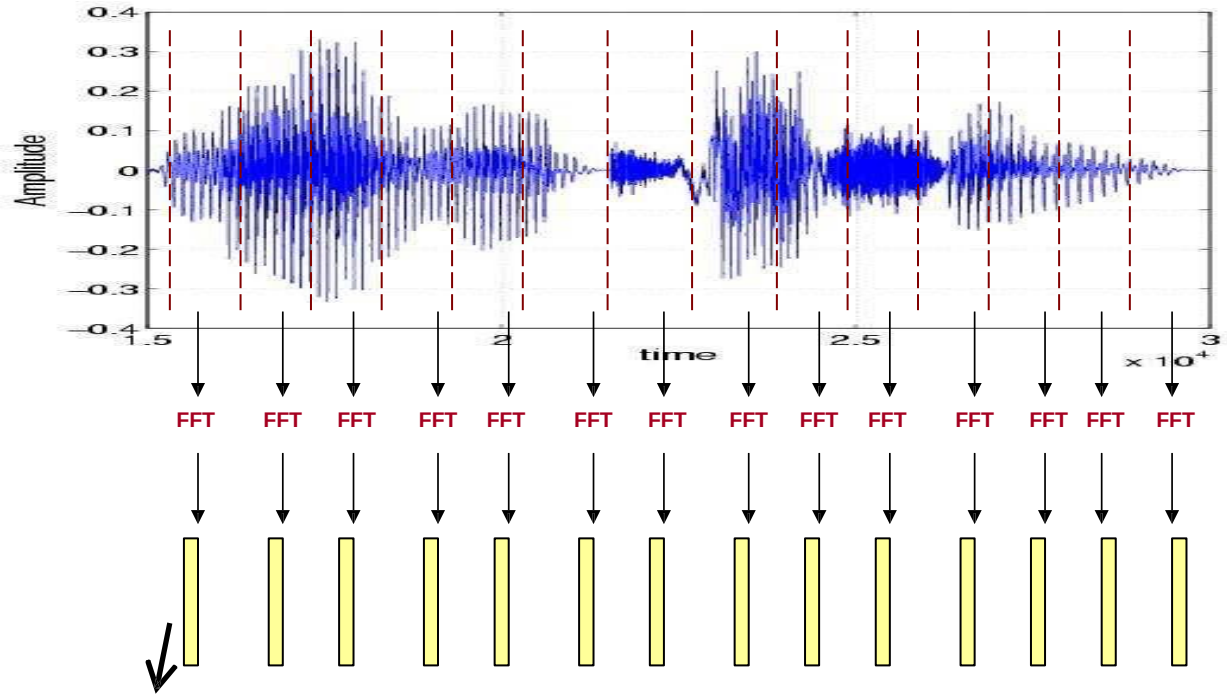


从语音信号到频谱向量的序列

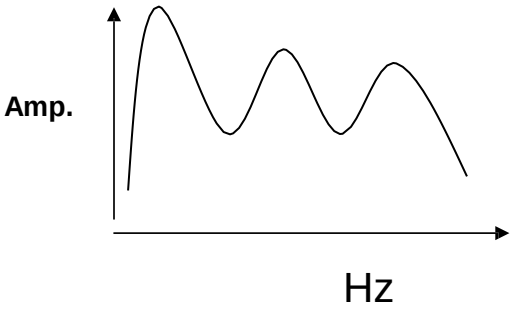


频谱

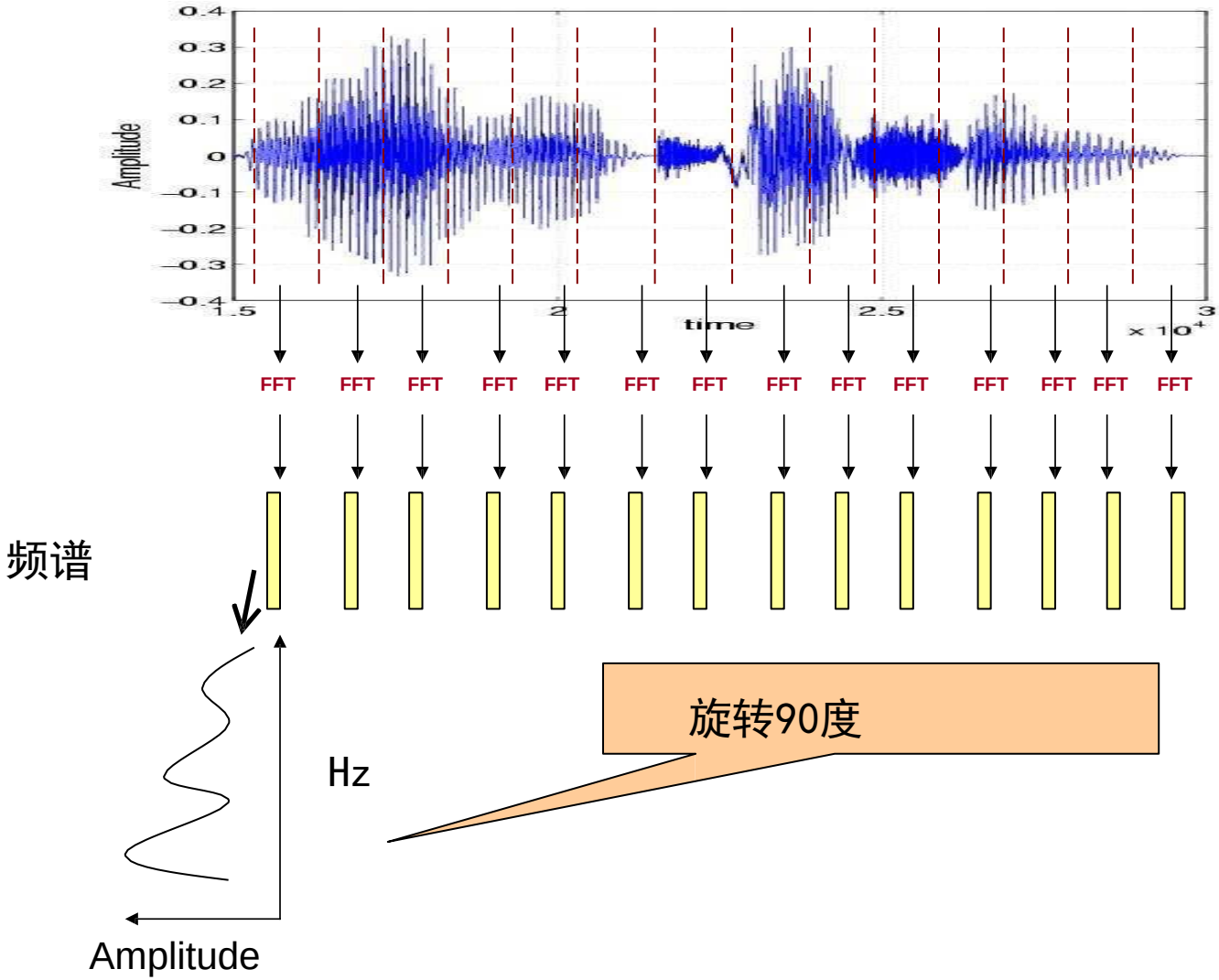
从语音信号到频谱向量的序列



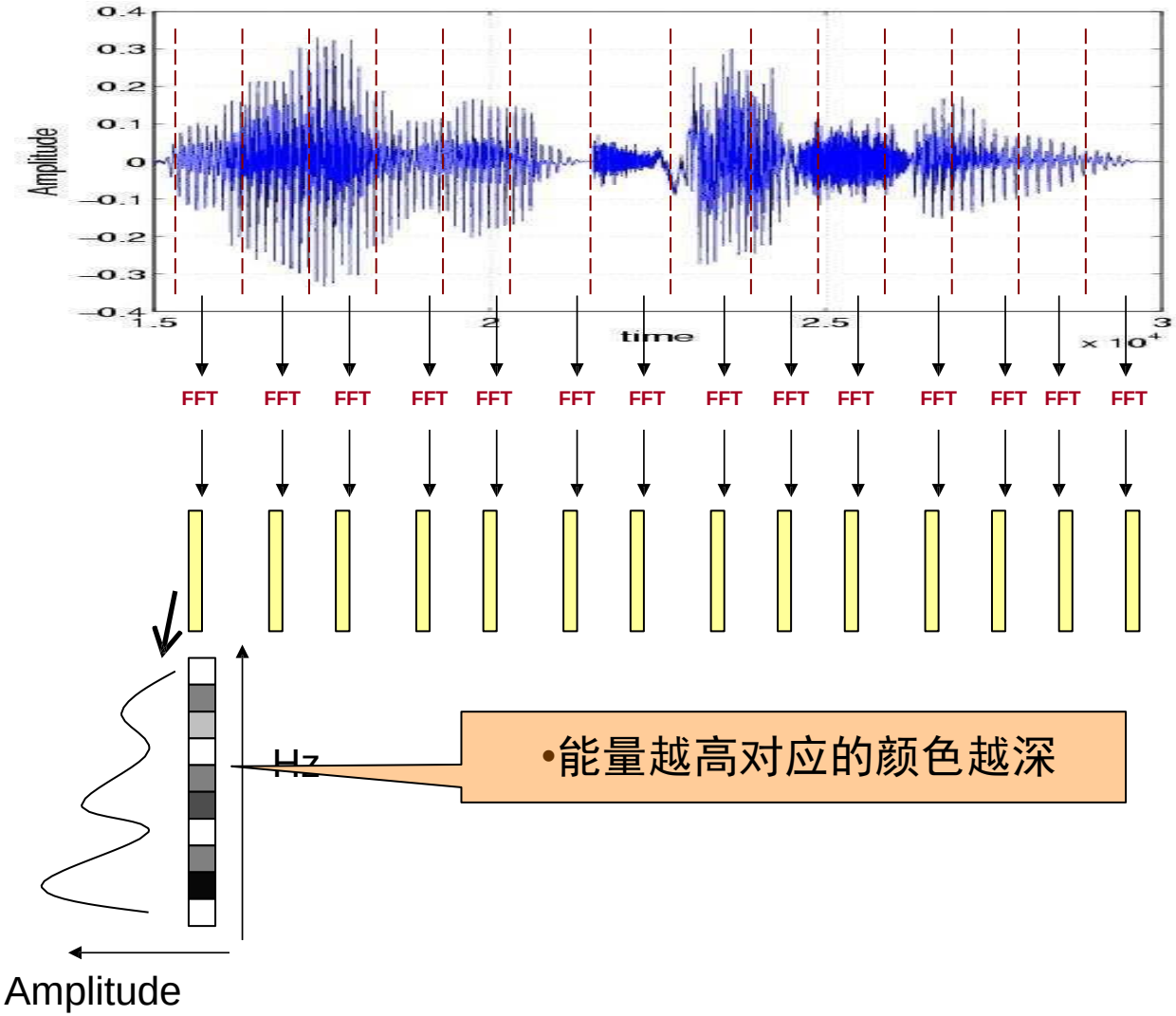
频谱



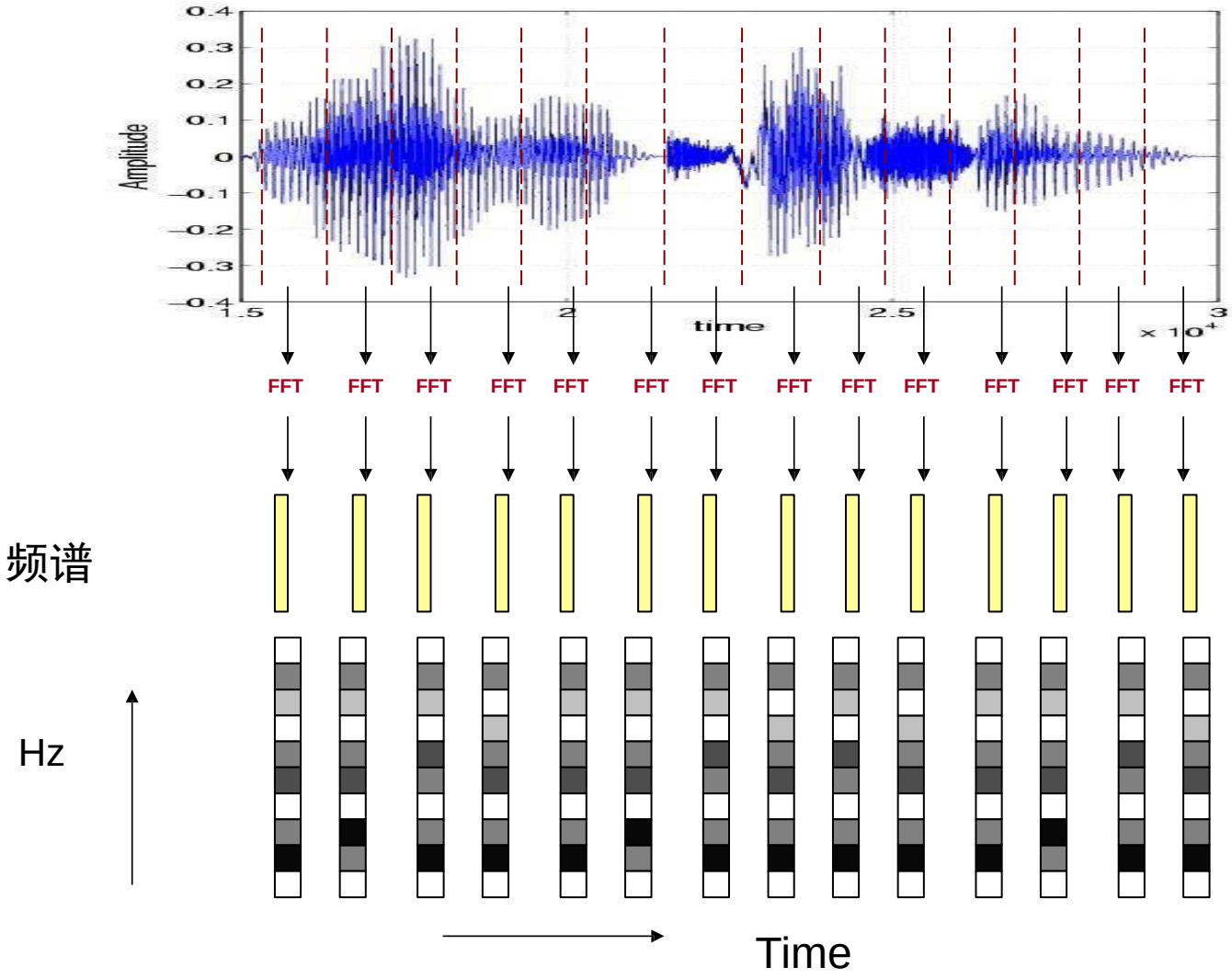
从语音信号到频谱向量的序列



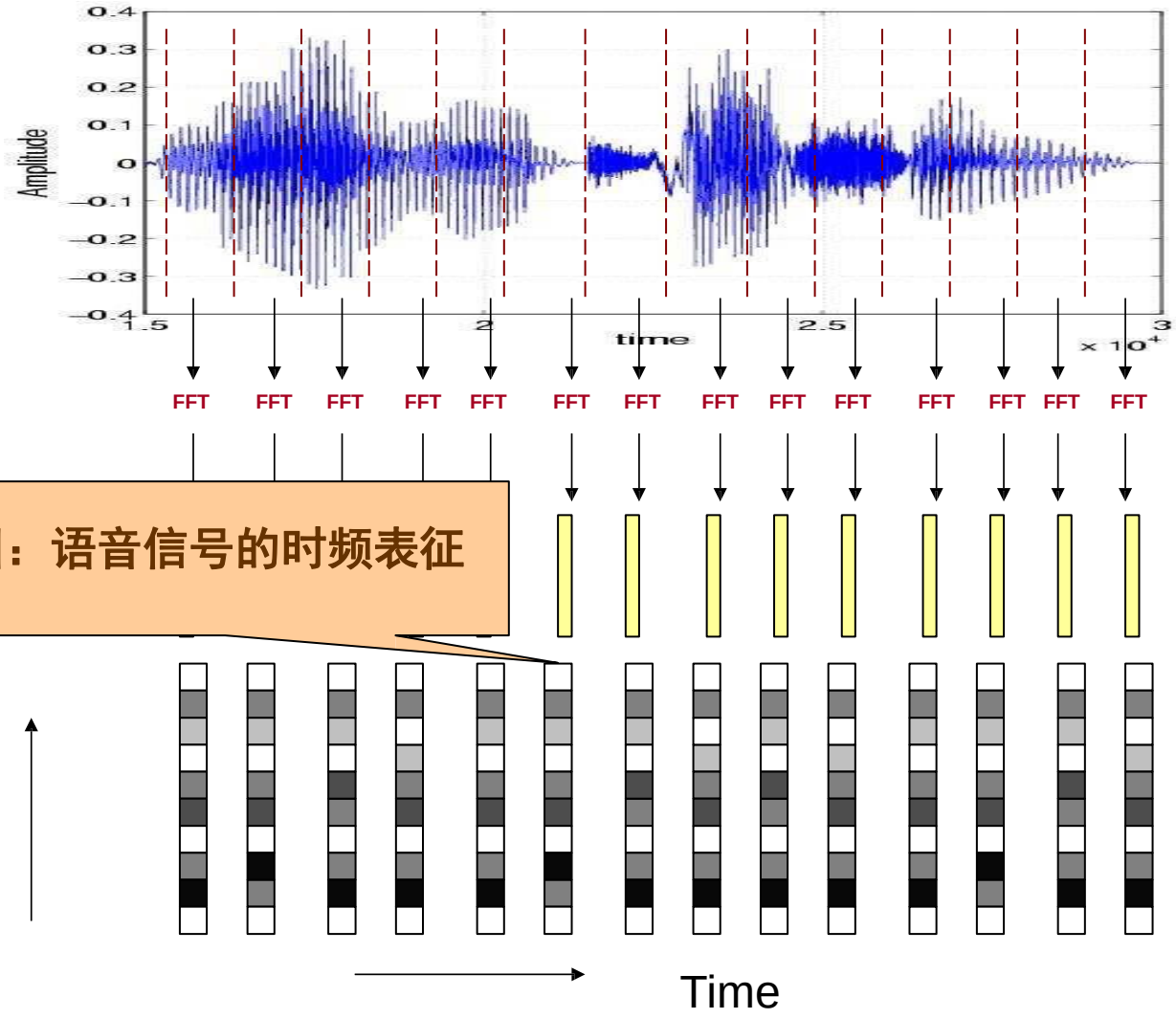
从语音信号到频谱向量的序列



从语音信号到频谱向量的序列

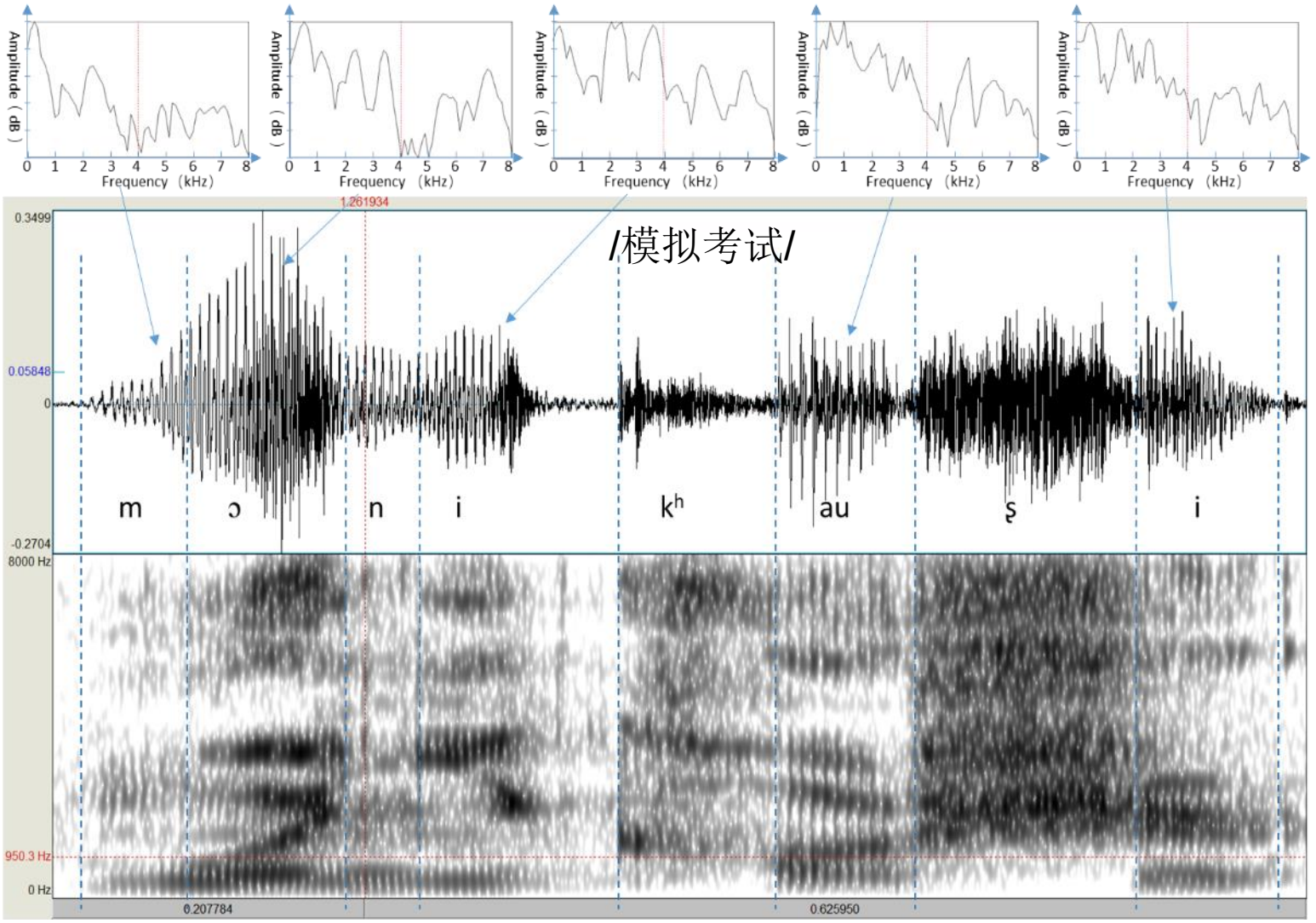


语谱图

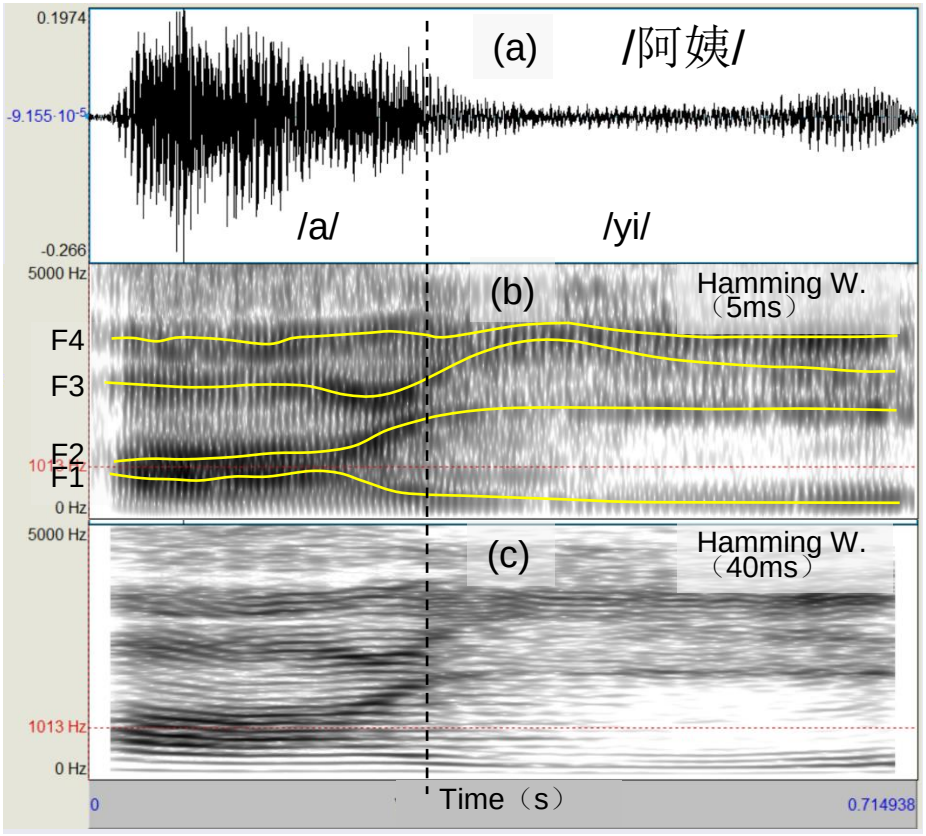


语谱图：语音信号的时频表征

语谱图的例子（了解）


Amplitude (dB)
Frequency (kHz)
1.261934
/模拟考试/
0.3499
0.05848
0
-0.2704
8000 Hz
950.3 Hz
0 Hz
m
ɔ
n
i
kʰ
au
ʂ
i
0.207784
0.625950

宽带语谱图与窄带语谱图



(b) 宽带 vs. (c)窄带语谱图

宽带滤波器(200Hz，窗长5ms)
—更低的频率分辨率
—更高的时间分辨率（周期性清晰）
观测共振峰(F1，F2，…)

窄带滤波器(25Hz，窗长40ms)
—更高的频率分辨率
—更低的时间分辨率（周期性被抹杀）
观测谐波(H1，H2，…)

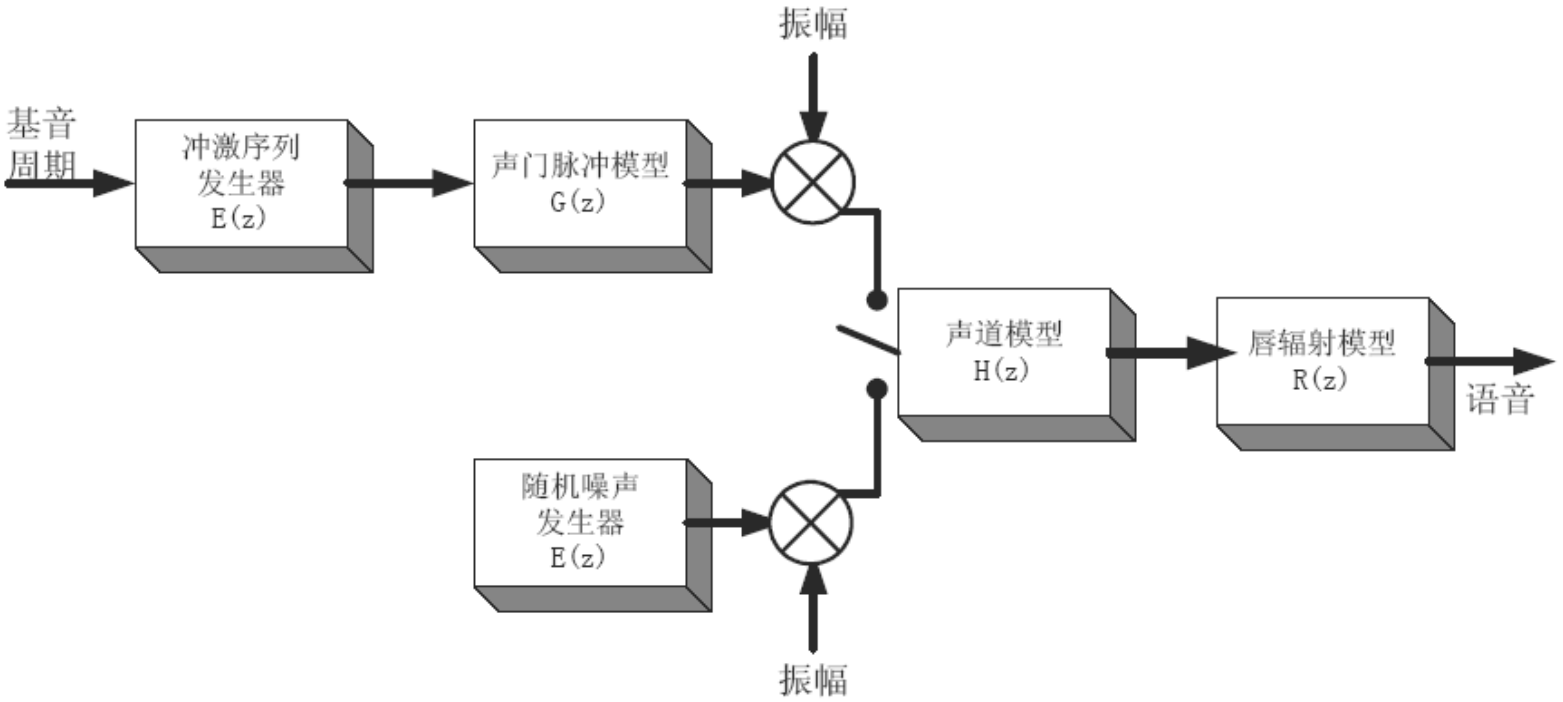
宽带：垂直条纹（垂直条纹的间隔时间为基音周期）
窄带：水平条纹（可以看见元音的共振峰频率及其随时间的变化），基音频率

第二章　语音信号处理的基础

- 3.1 语音信号的数字化和时频分析
- 3.2 语音产生与感知的数学模型
  - 3.2.1 语音产生的数学模型（**基本概念**）
  - 3.2.2 语音感知的数学模型（**了解**）
- 3.3 基于语音产生机理的特征分析方法
- 3.4 基于语音感知机理的特征分析方法

语音信号产生系统的线性模型

- 激励模型
- 声道模型
- 辐射模型

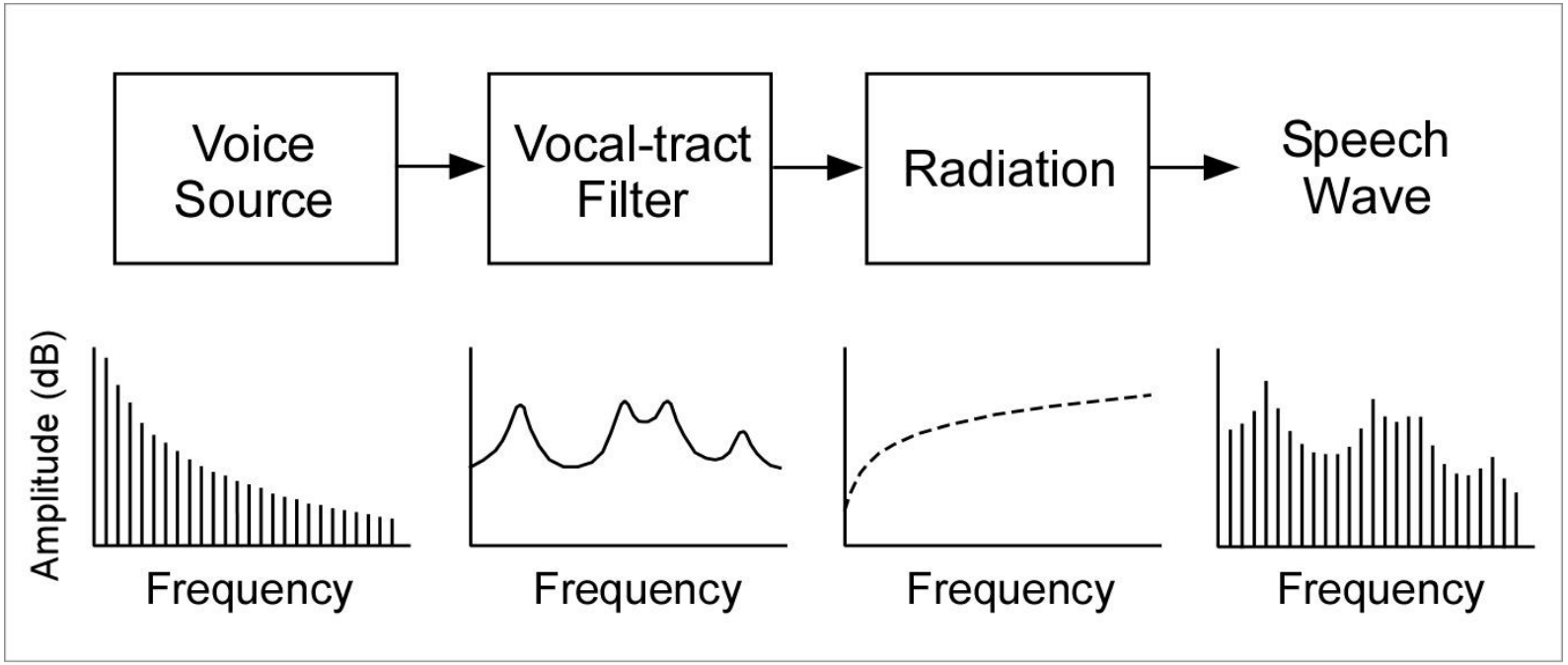


激励模型

- 激励模型非常复杂
- 声门脉冲模型
  - 浊音
    - 声带振动，声门脉冲
    - 如：斜三角形脉冲串
- 随机白噪声
  - 清音
    - 声带不振动，随机白噪声

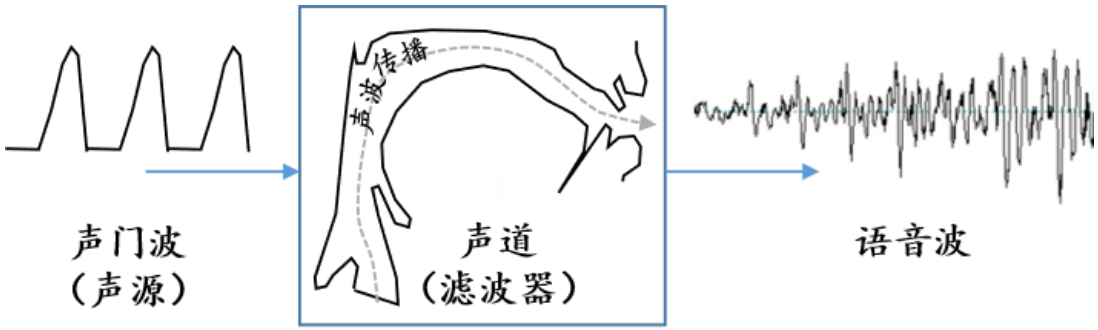
语音生成的声学理论

语音生成的声学理论 (Mueller, Chiba-Kajiyama, Fant)

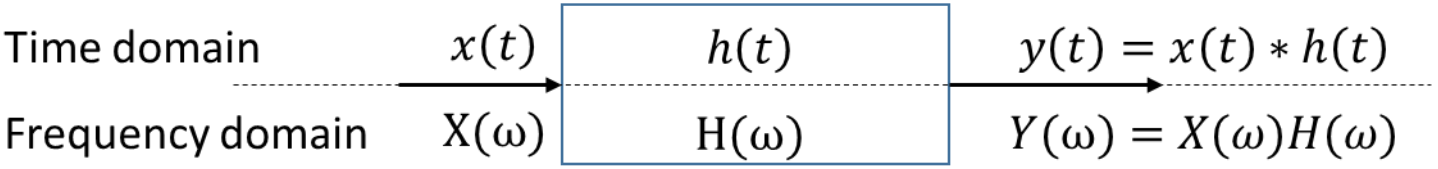


- 声带的振动产生了元音的声源。准周期脉冲通过声道，声道相当于滤波器，抑制某些频率，增强另外一些频率。其结果形成我们可以辨识的元音特征。
- 辅音声源处在声道中的不同位置，将声道分为前后腔。因此，声道需要两个以上的滤波器建模。

语音产生的声源-滤波器模型



- 发音器官连续运动形成时变的声道形状，由声源驱动产生语音信号。
- 发音器官的运动频率大约在50Hz以下，所以可以近似地认为声道形状在20 ms的范围内是稳定不变的
- 因此，语音生成系统在20-30ms的区间内可以用下面的线性时不变系统近似。

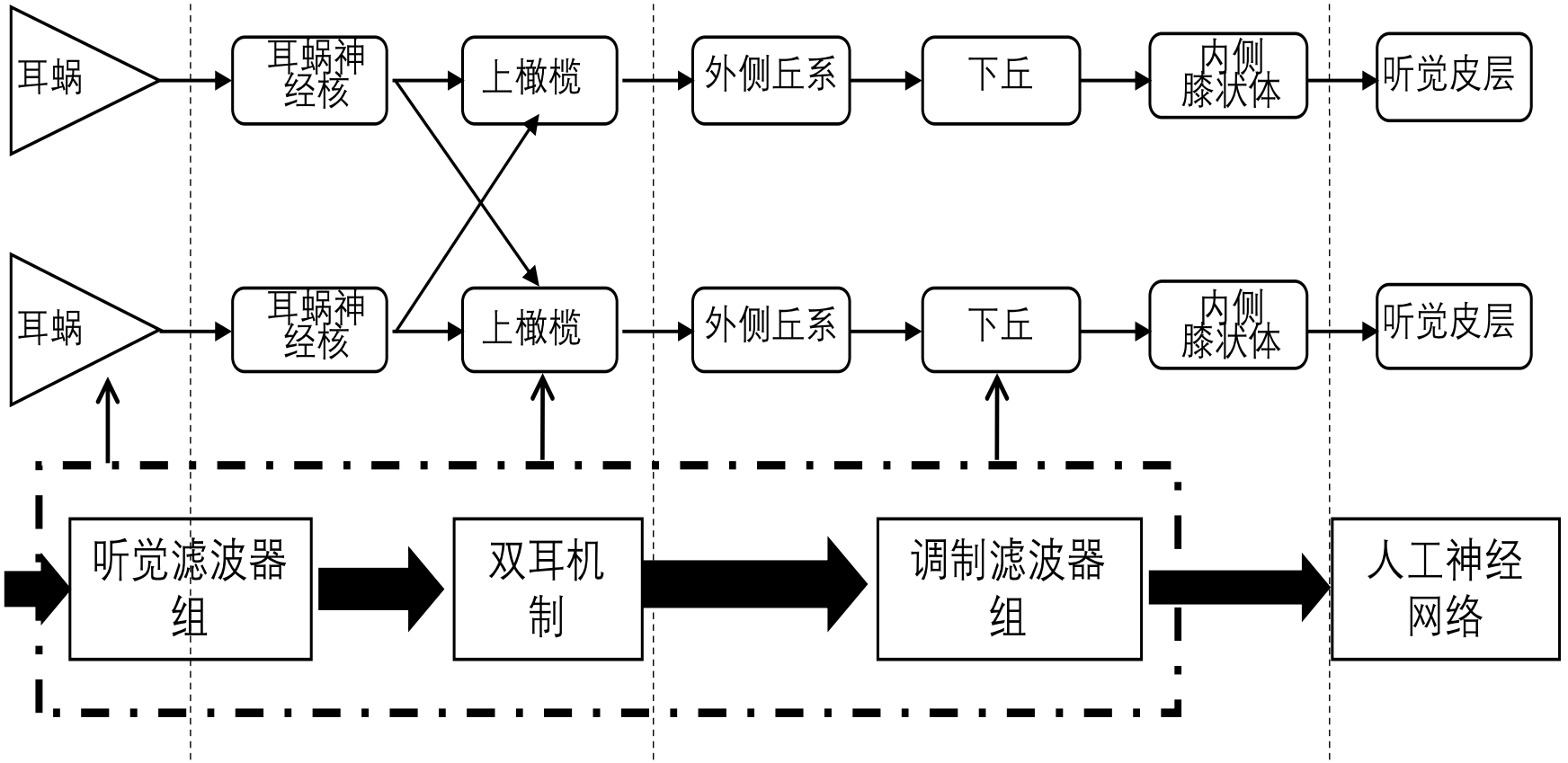


$$y(t) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)d\tau$$

第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
- 2.2 语音产生与感知的数学模型
  - 2.2.1 语音产生的数学模型
  - 2.2.2 语音感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
- 2.4 基于语音感知机理的特征分析方法

基于神经心理学研究的听觉数学模型



语音信号在听觉末梢系统中的声学频率分析是通过带通滤波器实现的。该带通滤波器被称为听觉滤波器。

听觉滤波机理与建模（基本概念）

- 听觉滤波器（带通滤波器）：实现听觉末梢系统中的声学频率分析。
- 等效矩形带宽(ERB)：根据心理物理实验结果获得的滤波器的带宽。
- ERB_N：听力正常人的ERB。ERB_N传递的能量与其对应的听觉滤波器相同，并显示它如何随着输入频率改变。

$$ERB_N = 24.7(4.37F/1000 + 1)$$

其中，F 是以Hz为单位的中心频率。

听觉滤波机理与建模（续）

伽玛通（gammatone）滤波器：基于听觉神经激活率与输入语音的相关性推导的冲激响应，得到冲激响应中心频率附近的幅度频率特性。Gammatone滤波器可以用一个Gamma分布和一个余弦信号的乘积来近似：

$$g_i(t) = at^{n-1}e^{-2\pi B(f_i)t}\cos(2\pi f_i t + \varphi)$$

$$B(f_i) = 0.1039 f_i + 24.7$$

a：调节比例的常数；

n：滤波器级数，一般取4；

$B(f_i)$：衰减速度，取值为正数，值越大衰减越快，脉冲响应长度越短；

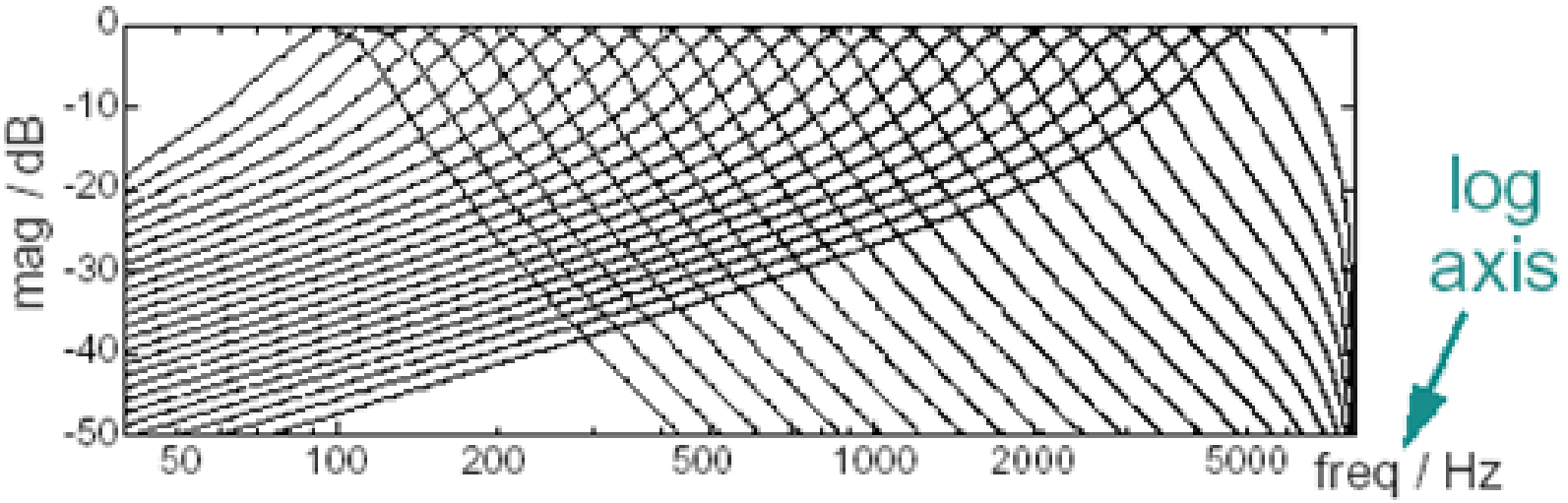
f_i：中心频率；

φ：相位，由于人耳对相位不敏感，可以省略；

t：时间，单位为秒。

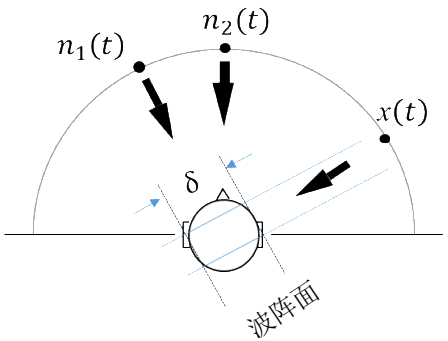
听觉滤波机理与建模（续）

- 伽马通（Gammatone）滤波器的频率响应

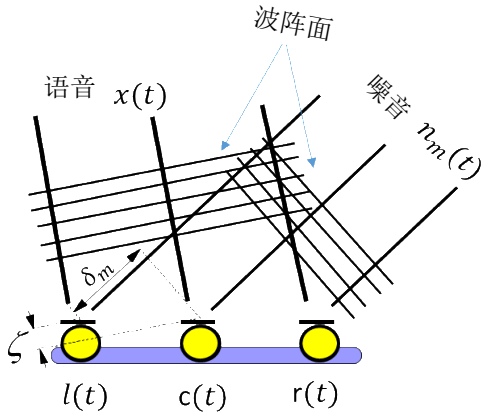


双耳声源定位及麦克风阵列

- 声音定位是听者在方向和距离上识别探测到的声音的位置或来源的能力，这是为了防御敌人的本能之一。
- 两耳时间差 （Interaural Time Difference：ITD）：比如，来自右侧的声音到达右耳的时间比到达左耳的时间要早。它适用于低频<800Hz。
- 两耳声强差 （Interaural Intensity Difference：IID）：比如，来自右边的声音在右耳比在左耳有更高的强度水平，因为头部遮挡了左耳。这些声强差异与频率密切相关，并且随着频率的增加而增加。它适用于高频>1600Hz。



(a)



(b)

$$x_1(t) = h_1(t) * s(t) + n_1(t)$$

$$x_2(t) = h_2(t) * s(t-\delta) + n_2(t)$$

第二章　语音信号处理的基础

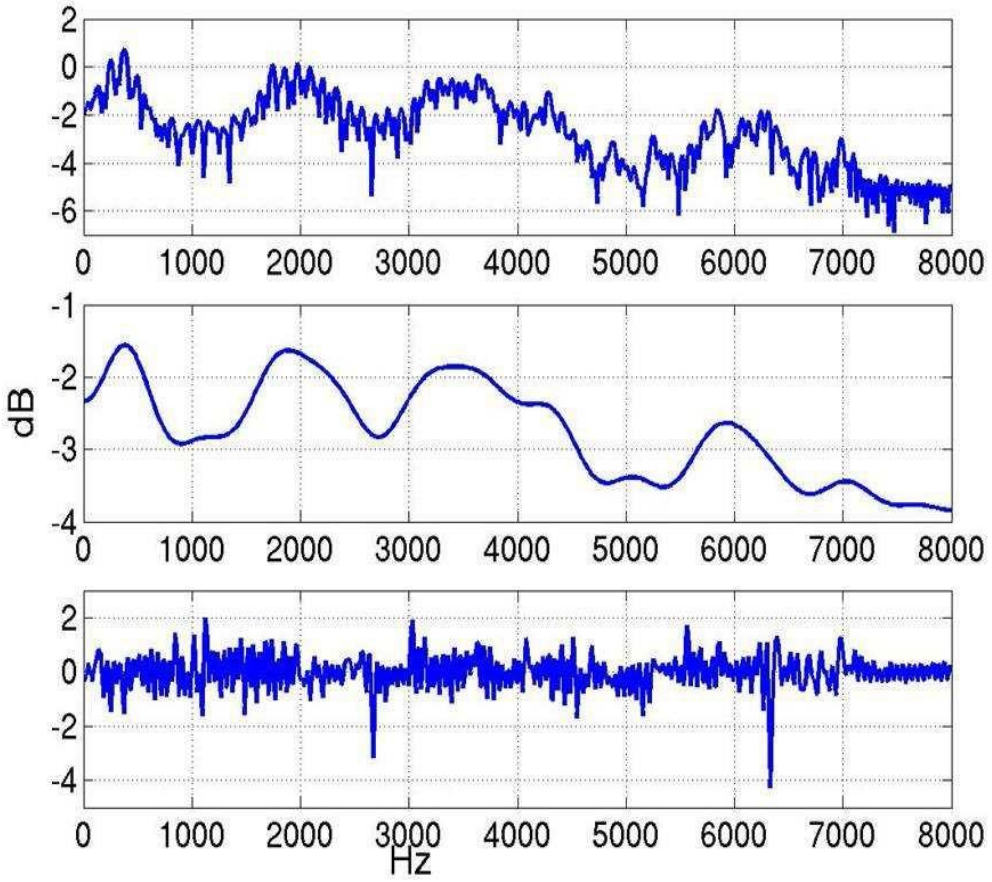
- 2.1 语音信号的数字化和时频分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
  - 2.3.1 倒谱分析（**掌握**）
  - 2.3.2 线性预测编码（**掌握**）
  - 2.3.3 语音基频的提取（**了解**）
- 2.4 基于语音感知机理的特征分析方法

语音频谱与语音包络分析

频谱

频谱包络

频谱细节



倒谱（Cepstrum）的定义

- 倒谱：信号的傅里叶变换经对数运算后再进行傅里叶反变换得到的谱。在语音信号处理中一般对频谱进行取幅度值运算。

  - 倒谱的定义

$$c_n = \mathrm{idft}(\log|\mathrm{dft}(x[n])|)$$

基于语音生成的分析方法-倒频谱分析

语音信号序列$x(n)$是声源$s(n)$和声道脉冲$h(n)$的卷积：

$$x(n) = \sum_{i=-\infty}^{n} s(i)h(n-i)$$

在频域上，语音信号的功率谱$X(k)$可以被描述为声源频谱$S(k)$和声道频谱$H(k)$的乘积

$$X(k) = S(k)H(k)$$

将上述公式取对数，我们可以得到

$$log|X(k)| = log|S(k)| + log|H(k)|$$

进行IFFT得：

$$C(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|X(k)|\, e^{j\omega n} d\omega$$
$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} (log|S(k)| + log|H(k)|)\, e^{j\omega n} d\omega$$

基于语音生成的方法-倒频谱分析

- 在IFFT之后，结果被转移到一个所谓的“类频率”（quefrency）空间而非时域空间。$C(n)$被称为“倒频谱”（cepstrum）

- 由于声源频谱$S(k)$的频率比声道频谱$H(k)$的频率高得多，

- $C(n)$（$n \geq F0$）对应声源信息。
- $C(n)$（$n < F0$）对应系统信息。

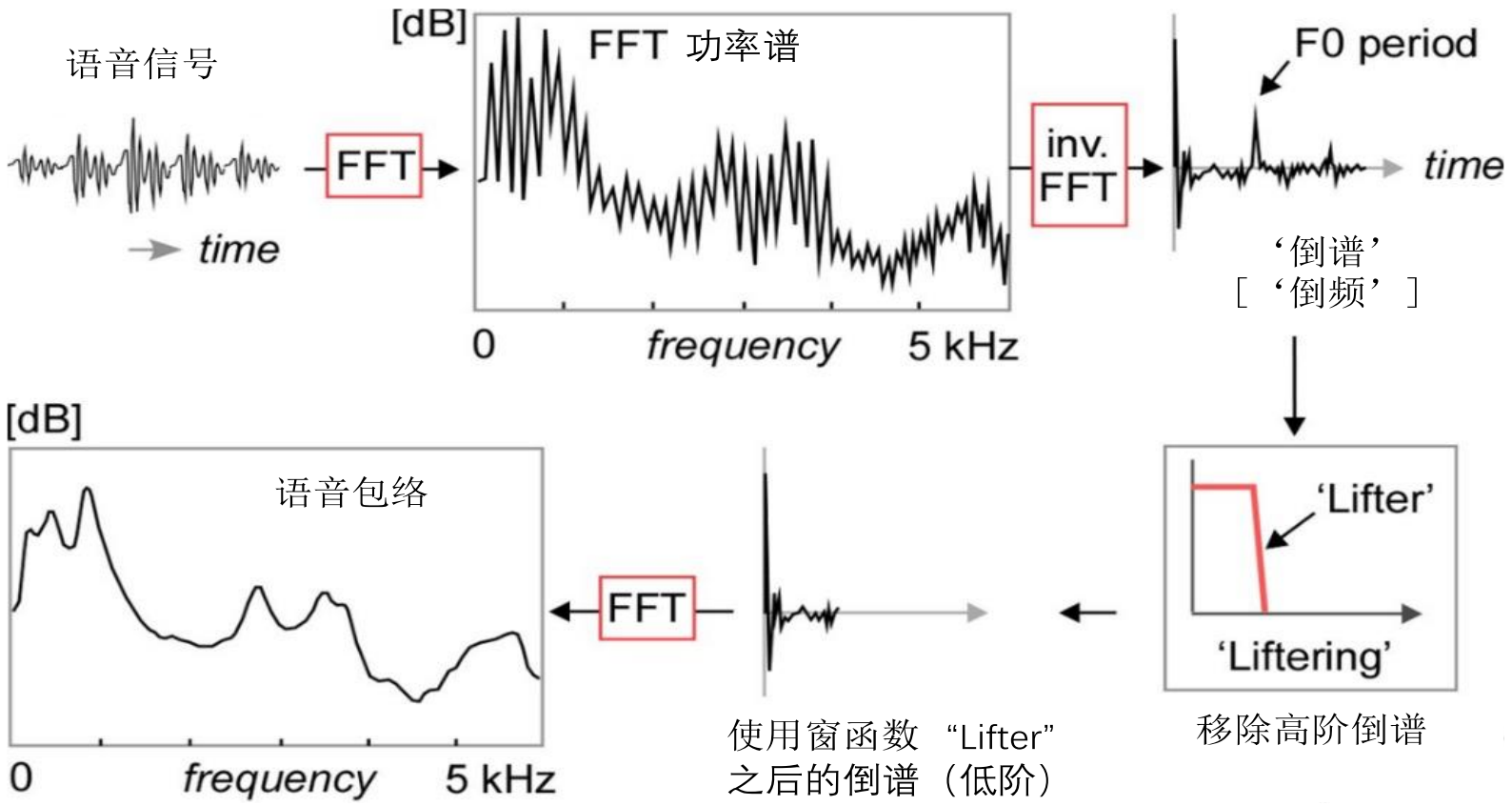
- 可以用一个“Lifter” 窗函数提取想要的信息，

$$C(n) = \begin{cases} C_L(n) & n < F0 \\ C_H(n) & n \geq F0 \end{cases}$$

基于FFT的倒频谱分析

‘倒频谱’ ← ‘频谱’

‘cepstrum’ ← ‘spectrum’

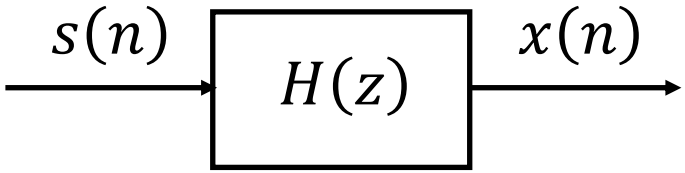


Oppenheim, A. V. & Shafer, R. W. (2004) From frequency to quefrency: A history of the cepstrum. IEEE Signal Processing Magazine, 21: 95-106.

第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
  - 2.3.1 倒谱分析
  - 2.3.2 线性预测编码
  - 2.3.3 语音基频的提取
- 2.4 基于语音感知机理的特征分析方法

线性预测的基本思路



语音信号产生的模型化

- 可以将语音$x(n)$看作是由输入序列$s(n)$（声源）激励一个全极点的系统$H(z)$而产生的输出。

- 自回归（auto-regressive）模型（简称AR模型）的传递函数为：

$$H(z)=\frac{G}{1-\sum_{i=1}^{p}a_i z^{-i}}$$

G为常数；为实数；为模型的阶数。

- 用系数$\{a_i\}$可以定义一个p阶线性预测器。

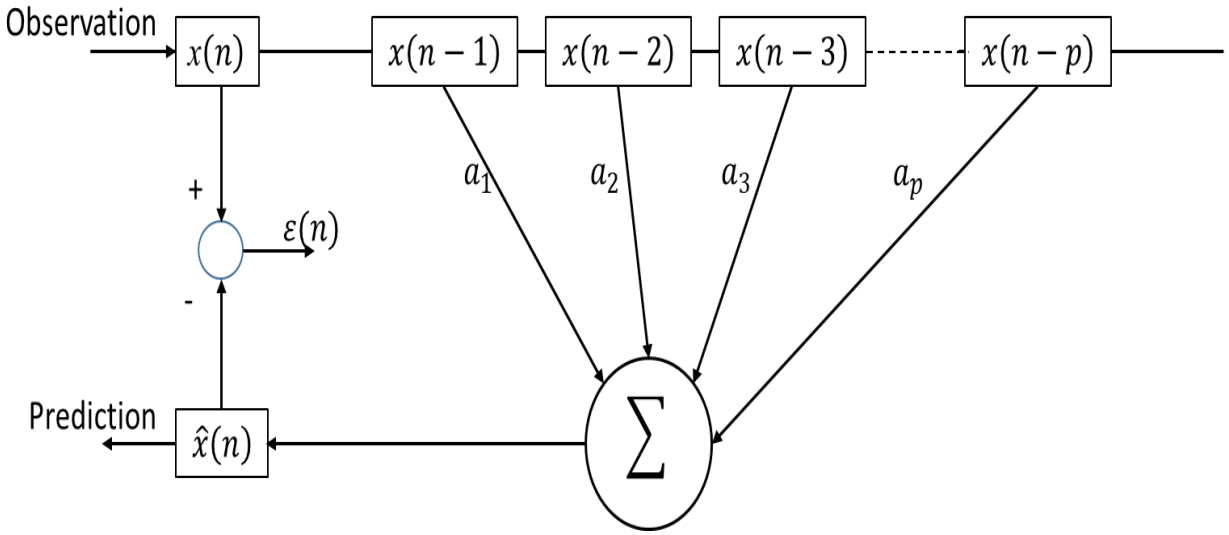
线性预测编码（LPC）

线性预测编码 （Linear predictive coding：LPC）是通过已知p个采样$x(n-i),\ \{1,2,\ldots,p\}$的线性组合推测信号的当前值 $x(n)$ ，

$$\hat{x}(n)=\sum_{i=1}^{p}a_i x(n-i)$$

$$\varepsilon(n)=x(n)-\hat{x}(n)=x(n)-\sum_{i=1}^{p}a_i x(n-i)$$

其中，$\varepsilon(n)$ 是观测值$x(n)$和预测值$\hat{x}(n)$的残差，$a_i\ (i=1,2,\ldots,\mathrm{p})$ 是加权系数。



线性预测编码(LPC)（续）（了解）

通过最小化均方根 $E[\varepsilon^2(n)]$使得 a_i最优化，

$$E[\varepsilon^2(n)] = E\left[\left[x(n) - \sum_{i=1}^{p} a_i x(n-i)\right]^2\right]$$

令 $$\frac{\partial E[\varepsilon^2(n)]}{\partial a_j} = -2E[\varepsilon(n)x(n-j)] = 0, \quad \{1 \le j \le p\}$$

$$E\left[x(n)x(n-j) - \sum_{i=1}^{p} a_i x(n-i)\, x(n-j)\right]$$
$$= r(j) - \sum_{i=1}^{p} a_i r(j-i) = 0, \qquad \{1 \le j \le p\}$$

其中， $r(j) = E[x(n)x(n-j)]$ 是 $x(n)$ 的自相关函数。

设 $\boldsymbol{r} = [r(1), r(2), \ldots r(p)]^{\boldsymbol{T}}$， $\boldsymbol{A} = \left[a_1, a_2, \ldots a_p\right]^{\boldsymbol{T}}$，且

$$\boldsymbol{R} = \begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{bmatrix}$$

线性预测编码(LPC)（续）（了解）

上式可以改写为联立方程

$$\boldsymbol{r} - \boldsymbol{RA} = 0$$

其中，$\boldsymbol{r}$ 是自相关向量，$\boldsymbol{R}$ 是自相关矩阵，而 $\boldsymbol{A}$ 为参数向量。该式被称为尤尔-沃克方程（Yule Walker Equation)。

基此，p个预测器的系数a_i可以通过尤尔-沃克方程获得。利用 $a_i\{i = 1,2,\ldots,\mathrm{p}\}$ 使得$E[\varepsilon^2(n)]$ 最小。令最小化的 $E[\varepsilon^2(n)]$ 为 E_{pm} ，则

$$E_{pm} = E[\varepsilon^2(n)]_{min} = r(0) - \sum_{i=1}^{p} a_i r(i)$$

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(p) \\ r(1) & r(0) & \cdots & r(p-1) \\ \vdots & \vdots & \ddots & \vdots \\ r(p) & r(p-2) & \cdots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ -a_1 \\ \vdots \\ -a_p \end{bmatrix} = \begin{bmatrix} E_{pm} \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

联立方程可用莱文逊-杜宾递推算法或者舒尔递推算法求解。

第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
  - 2.3.1 倒谱分析
  - 2.3.2 线性预测编码
  - 2.3.3 语音基频的提取
- 2.4 基于语音感知机理的特征分析方法

时域基频(F0)的抽取方法

计算语音的基频可以在时域、频域或两者上进行。

- 基于自相关的方法：使用低通滤波器(LPF)在前两个谐波上截断频率，例如900Hz

$$R(m) = \frac{1}{N+1}\sum_{n=0}^{N-1-m} x(n)x(n+m),\ \{m = 0,1,\cdots,N-1\}$$

使用中心截波，找到 $R(0) > R(p) > R(2p)\cdots$ 的峰值。

- 基于LPC的方法：预测的残差与声源信号的周期有密切的关系

$$\varepsilon(n) = x(n) - \hat{x}(n) = x(n) - \sum_{i=1}^{p} a_i x(n-i)$$

残差信号的谱接近平坦并且共振峰的效应在残差信号中已被去除掉了。因此，利用预测残差作自相关分析，并在恰当的范围内测出最大峰值，进行基音检测，可获得比较理想的基频检测结果。

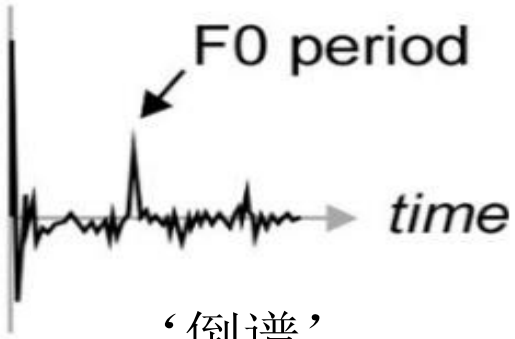
频域基频(F0)的抽取方法

- 基于倒频谱的方法：在“类频率”轴上分离声源和滤波器

$$C(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi} log|X(k)|\, e^{j\omega n}d\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}(log|S(k)|+log|H(k)|)\, e^{j\omega n}d\omega$$

$$C(n)=\begin{cases} C_L(n) & n<F0 \\ C_H(n) & n\geq F0 \end{cases}$$

通过“lifter”获得F0



‘倒谱’
[‘倒频’]

第二章　语音信号处理的基础

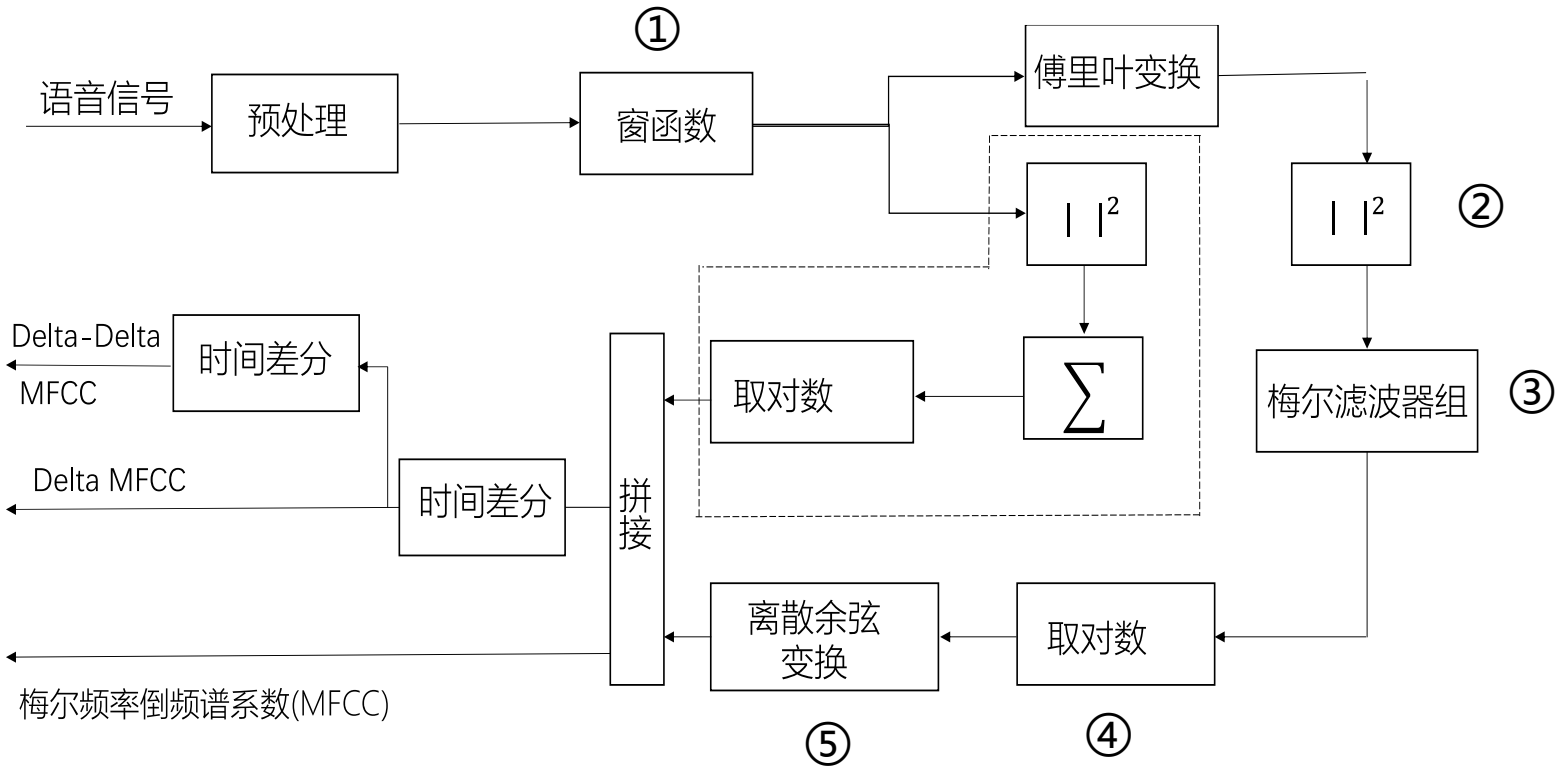
- 2.1 语音信号的数字化和时频分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
- 2.4 基于语音感知机理的特征分析方法
  - 2.4.1 梅尔频率倒谱系数(MFCC) （掌握）
  - 2.4.2 感知线性预测(PLP) （了解）

梅尔尺度（Mel-scale）

- 人耳对不同频率的灵敏度响应是不同的，即以赫兹为单位时，人耳的听觉系统是一个非线性的系统，整体成对数曲线趋势。
- 梅尔频率的产生很好地模拟了人耳的感知特性曲线。梅尔频率与客观音频率 f 的转换关系为：

$$\mathrm{Mel}(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

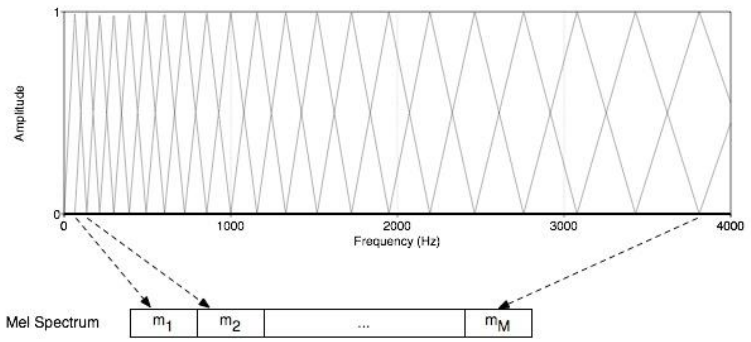
梅尔频率倒谱系数(MFCC)--- FFT



① 语音信号经过预加重、加窗、分帧的预处理过程，得到帧序列语音信号。

② 对帧序列语音信号进行离散傅里叶变换（DFT），取模的平方得到离散能量谱$|X(k)|^2, 1 \le k \le \frac{N}{2} - 1$。令$S(k) = |X(k)|^2$。

梅尔频率倒谱系数(MFCC)--- 梅尔滤波器组

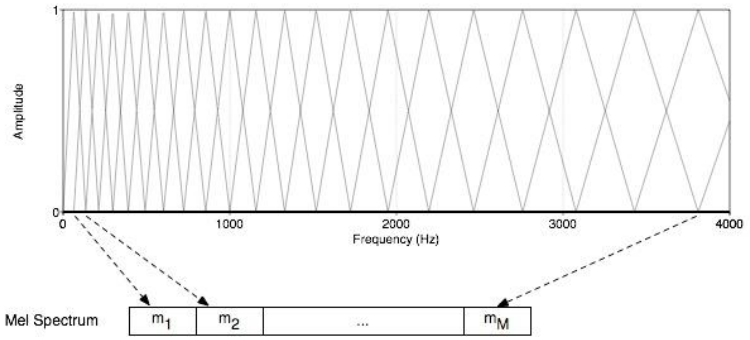


③ 计算 $S(k)$ 通过一组梅尔频率三角滤波器组得到的一组系数。采样频率为 f_s，则在$[0, Mel(f_s/2)]$范围内等间隔地选取 L 个频率点作为三角滤波器的中心频率$f_c(i)$,$1 \leqslant i \leqslant L$。第$i$个滤波器的低频边界和高频边界分别记为 $f_{\mathrm{l}}(i)$ 和 $f_{\mathrm{h}}(i)$。

$$f_c(i) = f_{\mathrm{l}}(i+1) = f_{\mathrm{h}}(i-1)$$

$S(k)$ 映射到梅尔频率域得 $S'(f_k)$

梅尔频率倒谱系数(MFCC)--- 梅尔滤波器组（续）

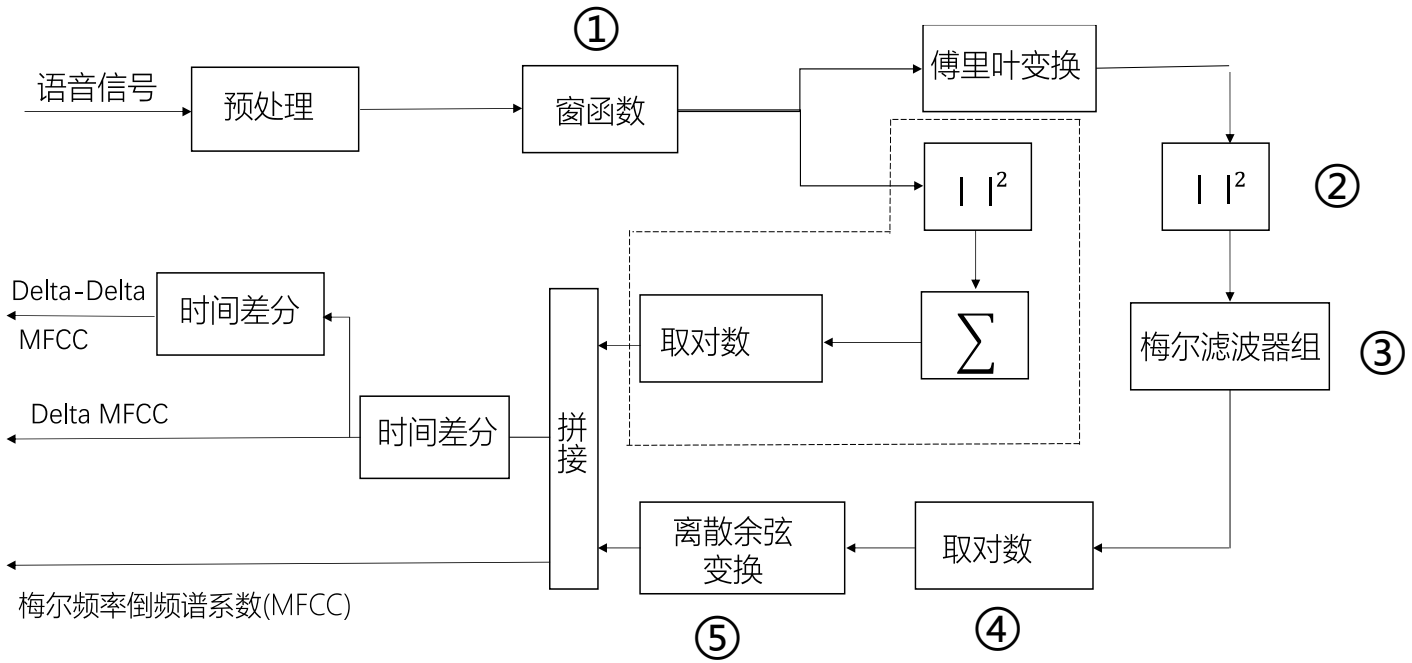


滤波器的输出为

$$m(i)=\sum_{k=1}^{n}W(k,i)\mid S'(f_k)\mid,\quad 1\leqslant i\leqslant L$$

$$W(k,i)=\begin{cases}\dfrac{f_k-f_1(i)}{f_c(i)-f_1(i)}, & f_1(i)\leqslant f_k\leqslant f_c(i)\\[2ex]\dfrac{f_n(i)-f_k}{f_n(i)-f_c(i)}, & f_c(i)<f_k\leqslant f_h(i)\end{cases}$$

梅尔频率倒谱系数(MFCC)--- DCT

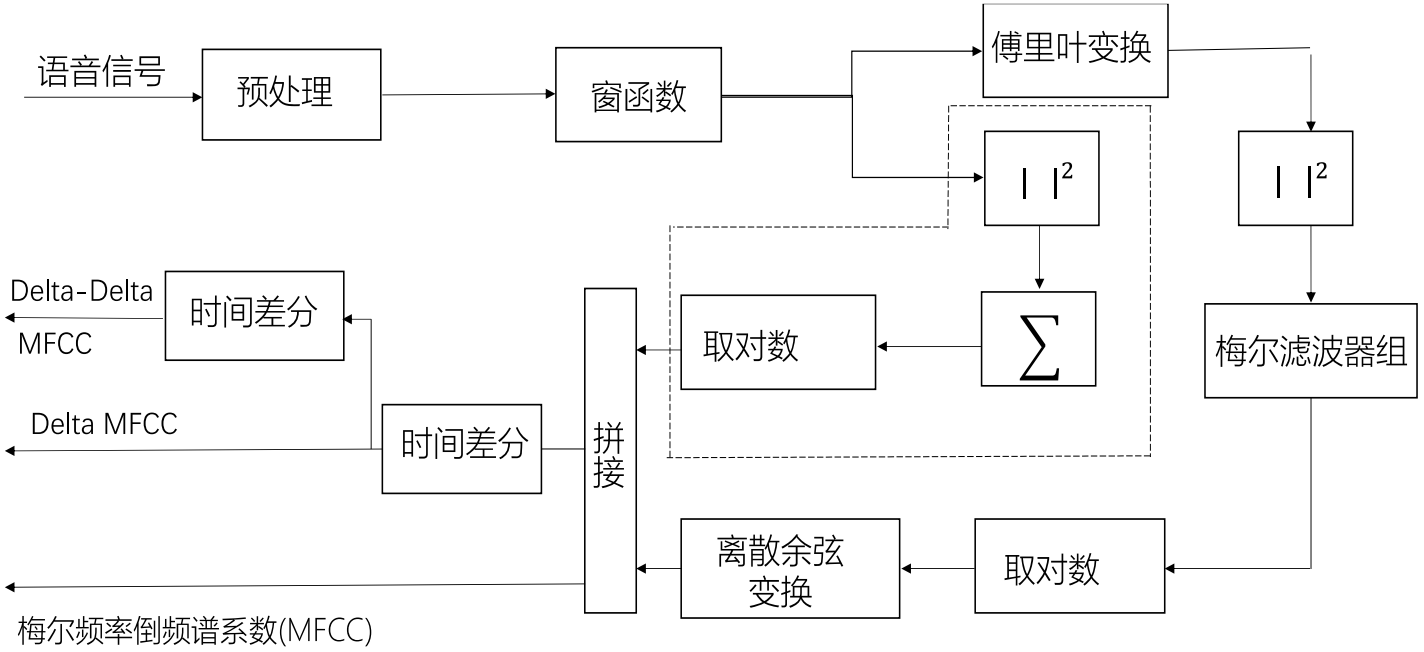


④ 对 $m(i)$ 取对数，得到 $\lg[m(i)]$。

⑤ 求 $\lg[m(i)]$ 的离散余弦变换(discrete cosine transform，DCT)，得到 MFCC 系数：

$$C_{MFCC}(n)=\sum_{i=1}^{L}\lg[m(i)]\cos\left[(i-0.5)\frac{n\pi}{L}\right]$$

MFCC的计算流程图及MFCC系数的物理意义

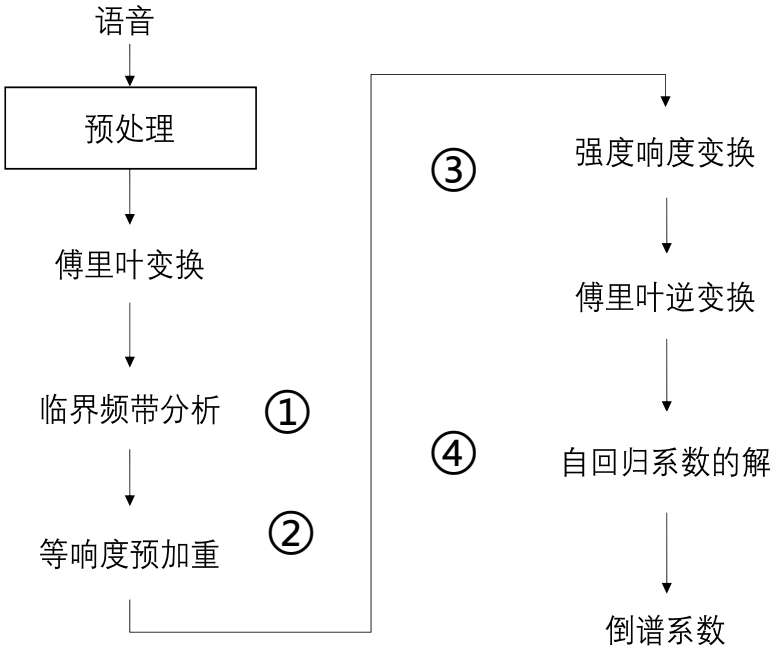


- C0：语音的平均强度，在语音识别中一般不直接使用。
- 随着阶数增加，MFCC表示频谱的更多细节。但是无论是MFCC还是LPC系数都不能直接展示共振峰的频谱包络的细节信息。

第二章　语音信号处理的基础

- 2.1 语音信号的数字化和时频分析
- 2.2 语音产生与感知的数学模型
- 2.3 基于语音产生机理的特征分析方法
- 2.4 基于语音感知机理的特征分析方法
  - 2.4.1 梅尔频率倒谱系数(MFCC)
  - 2.4.2 感知线性预测(PLP)

感知线性预测(PLP)的计算



①临界频带分析：

$$\Omega(\omega) = 6log\{\omega/1200\pi + [(\omega/1200\pi)^2 + 1]^{1/2}$$

将频谱P(ω)映射到Bark频率，得到N频带Ω_i

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega_i - \Omega)\psi(\Omega), \quad \{i = 1,2,\dots,N\}$$

②等响度预加重：

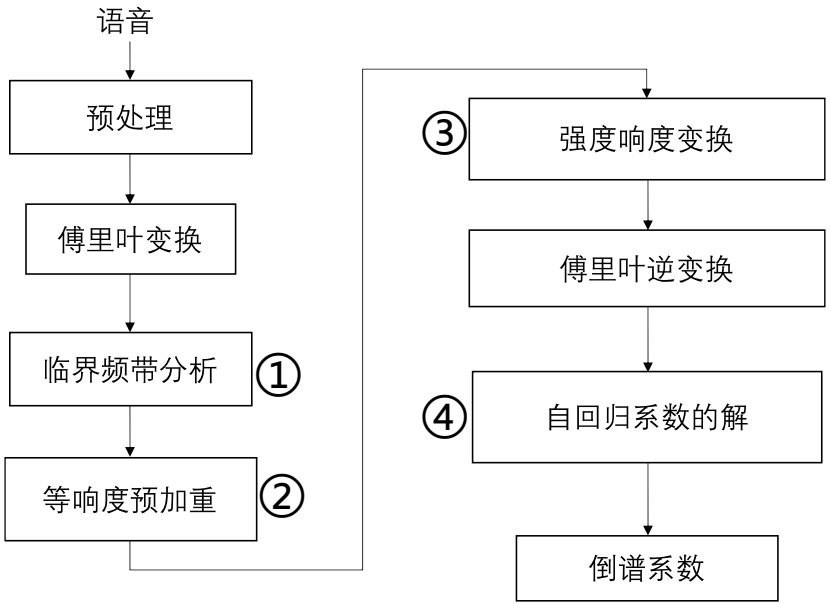
模拟人耳大约40dB等响曲线E(ω)对$\Theta(\Omega_i)$进行等响度曲线预加重，即

$$\Gamma(\Omega_i) = E(\omega_i)\Theta(\Omega_i), \quad \{i = 1,2,\dots,N\}$$

ω_i表示第i个临界带听觉谱的中心频率所对应的频率。其中

$$E(\omega_i) = [(\omega_i^2 + 56.8 \times 10^6)\omega_i^4]/[(\omega_i^2 + 6.3 \times 10^6)^2 \times (\omega_i^2 + 0.38 \times 10^9)]$$

感知线性预测(PLP)的计算（续）



③强度-响度转换：近似声音强度与人耳感知到的响度之间的非线性(1/3幂律听觉)

④计算倒频谱系数：在逆离散傅里叶变换(IDFT)后，计算12阶的LPC和16阶的倒频谱系数。最后的结果就是PLP特性参数。

PLP的一个重要优势就是它对阶数的敏感程度远小于LPC。

第一次上机实践课

内容：语音特征提取算法实现
时间：9月19日
机房：47教8号机房