语音识别：从入门到精通

第二讲：语音信号处理及特征提取

主讲人 孙思宁

博士，毕业于西北工业大学

ssning2013@gmail.com

内容提要

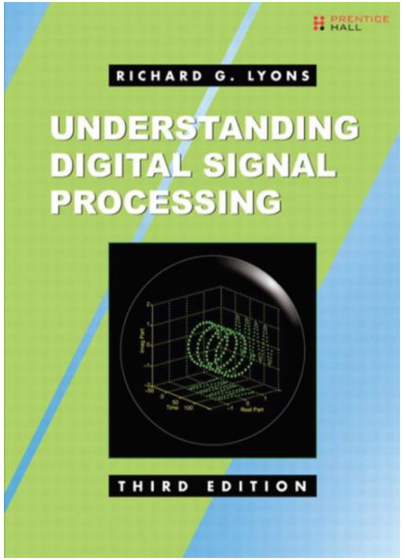
1.数字信号处理基础

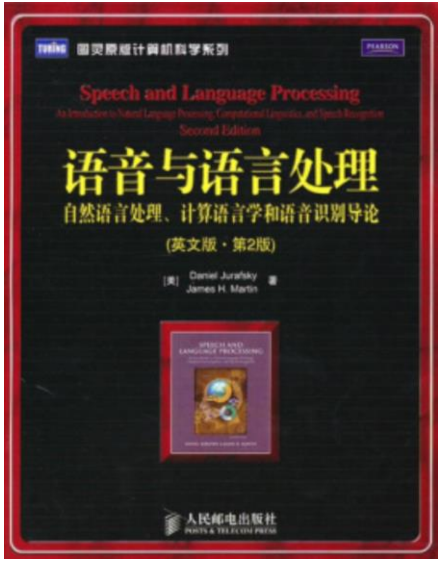
- 基础知识
- 傅里叶分析

2.常用特征提取

- 特征提取流程
- Fbank
- MFCC

3.课后实践



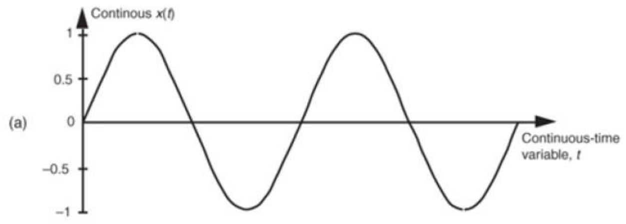


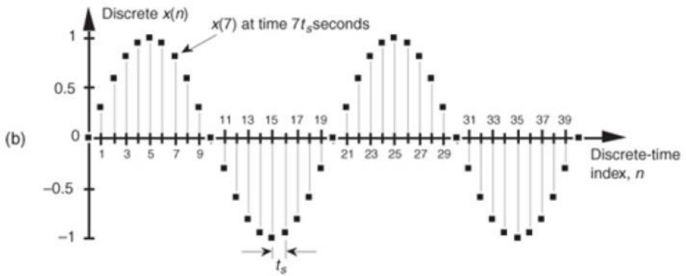
http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

数字信号处理基础

■ 模拟信号到数字信号转化（ADC）

■ 在科学和工程中，遇到的大多数信号都是连续的模拟信号，例如电压随着时间的变化，一天中温度的变化等等，而计算机只能处理离散的信号，因此，必须对这些连续的模拟信号进行转化，通过采样和量化，转换成数字信号。





数字信号处理基础

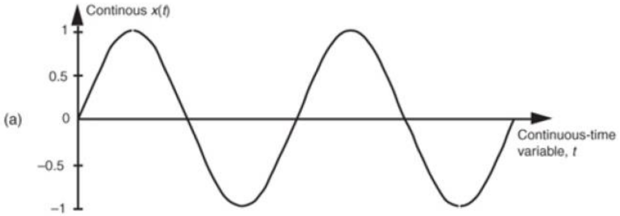
- 以正弦波为例，理解一些基础定义
  - 考虑一个正弦波（a）

$$x(t) = \sin(2\pi f_0 t)$$

其中f_0表示信号本身的频率，单位为Hz

如果我们对此正弦波进行采样，每隔t_s秒进行一次采样，并使用一定范围的离散数值表示采样值，则可以得到采样后的离散信号（b）

$$x(n) = \sin(2\pi f_0 n t_s)$$



数字信号处理基础

- 离散信号中的定义

$$x(n) = \sin(2\pi f_0 n t_s)$$

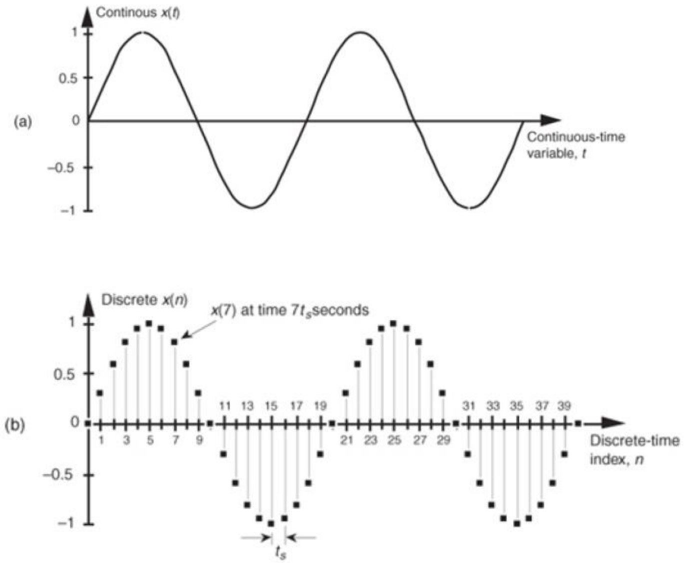
其中

t_s为采样周期;

$f_s = 1/t_s$,为采样频率，或采样率，表示1s内采样的点数;

$n = 0,1, \ldots \ldots$ 为离散整数序列

问题：如果给定一个正弦波采样后的序列，如（b），可以唯一的恢复出一个连续的正弦波吗?



数字信号处理基础

- 频率混叠

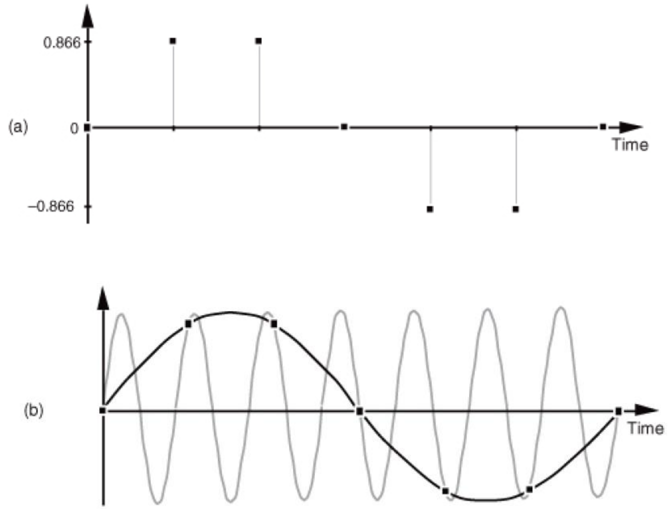
首先，请尝试使用图（a）中的采样点画出其对应的连续正弦波

图（b）给出了两种可能的画法，也就是说，不同频率的正弦波，经过采样后，完全有可能出现相同的离散信号！为什么？

$$
\begin{aligned}
x(n) &= \sin(2\pi f_0 n t_s) \\
&= \sin(2\pi f_0 n t_s + 2\pi m) \\
&= \sin(2\pi(f_0 + \frac{m}{n t_s}) n t_s)
\end{aligned}
$$

如果$m = kn$，k为整数（一般为常数），因为n为整数，m也必须为整数，若m=kn满足，则k必须为整数，对于任意n，m都为整数

$$X(n) = \sin(2\pi(f_0 + k f_s) n t_s)$$



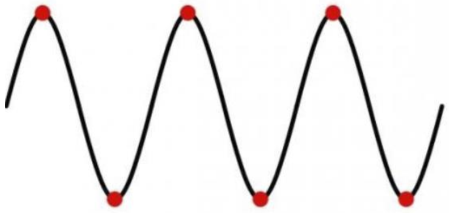
数字信号处理基础

- 奈奎斯特采样定律

采样频率大于信号中最大频率的两倍!

$$f_s/2 \geq f_{max}$$

即，在原始信号的一个周期内，至少要采样两个点，才能有效杜绝频率混叠问题。



问题1：如果对语音模拟信号进行采样率为16000Hz的采样，得到的离散信号中包含的最大频率是多少?

问题2：对一个采样率为16K的离散信号进行下采样，下采样到8K，为什么要需要首先进行低通滤波?

离散傅里叶变换

- 为什么要进行DFT?

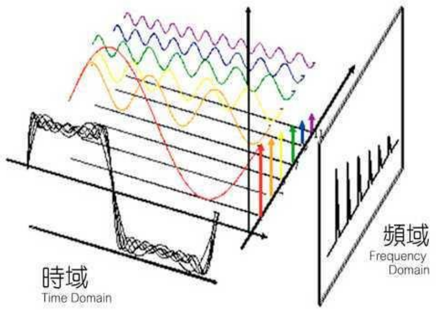
  DFT将时域信号变换到频域，分析信号中频率成分

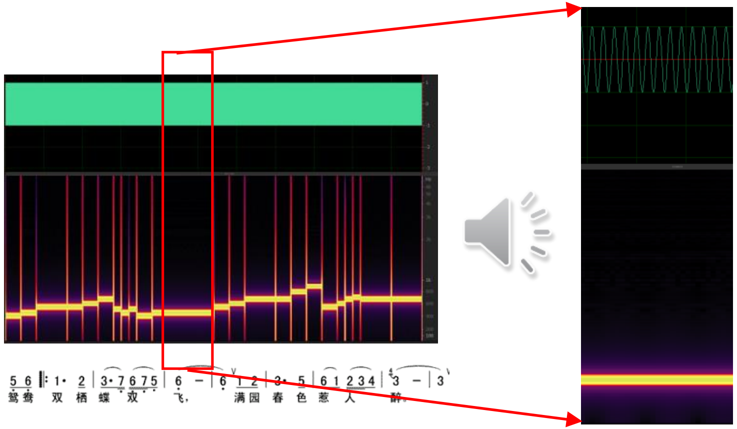
- 什么信号可以进行DFT?

  时域离散且周期的信号

- 非周期离散信号可以吗?

  需要进行周期延拓



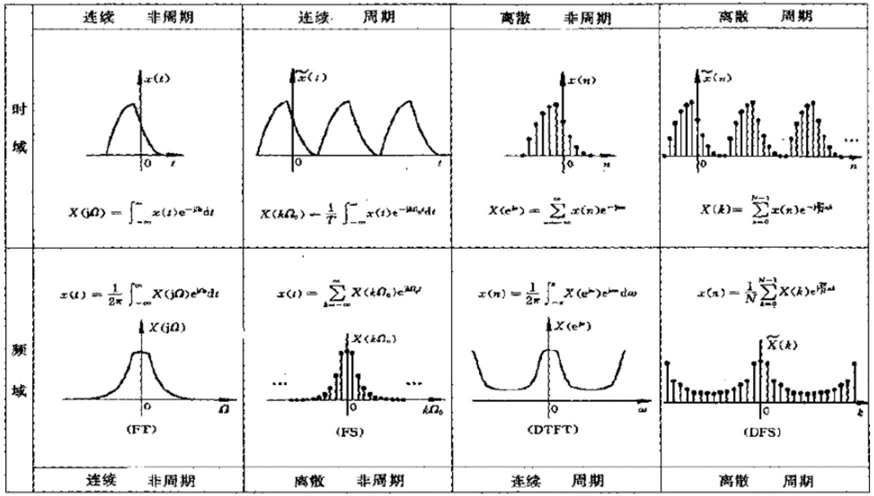


音名	频率/Hz
低音 1	261.63
低音 2	293.67
低音 3	329.63
低音 4	349.23
低音 5	391.99
低音 6	440
低音 7	493.88

离散傅里叶变换

- 傅里叶家族

傅里叶变换 | 傅里叶级数 | 离散时间傅里叶变换 | *离散傅里叶变换（DFT）*



只有DFT是在时域和频域上都具有离散和周期的特点，因此，也只有DFT可以用计算机来处理！

1. 时域上的采样（离散化），导致了频域上的周期，为什么？
2. 时域上的周期，导致了频域上的离散，为什么？

离散傅里叶变换

- DFT定义:给定一个长度为N的离散信号，DFT定义了对应的离散频域序列X(m)为:

$$X(m)=\sum_{n=0}^{N-1}x(n)e^{-j2\pi nm/N}$$

其中:

$j=\sqrt{-1},$

e为自然对数底

m频域序列的索引，

$m=0,1,2\ldots,N-1$

X(m)为DFT的第m个输出

- 根据欧拉公式，DFT的公式还可以为:

$$X(m)=\sum_{n=0}^{N-1}x(n)[\cos\left(\frac{2\pi nm}{N}\right)-j\sin(\frac{2\pi nm}{N})]$$

- DFT本质上是一个线性变换:

$$\begin{bmatrix} X(0) \\ X(1) \\ X(2) \\ \vdots \\ X(N-1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & e^{-j2\pi/N} & \cdots & e^{-j2\pi(N-1)/N} \\ 1 & e^{-j2\pi\cdot 2/N} & \cdots & e^{-j2\pi\cdot 2(N-1)/N} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & e^{-j2\pi\cdot(N-1)/N} & \cdots & e^{-j2\pi\cdot(N-1)(N-1)/N} \end{bmatrix} \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ \vdots \\ x(N-1) \end{bmatrix}$$

离散傅里叶变换

例题1：给定信号

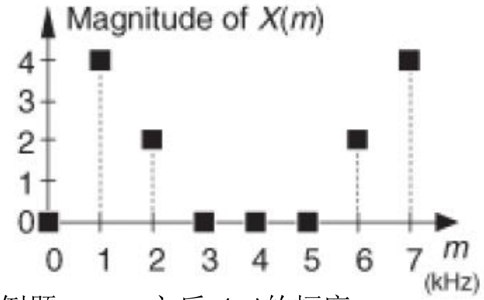
$x(n)=\sin(2\pi\cdot 1000\cdot nt_s)+0.5\sin\left(2\pi\cdot 2000\cdot nt_s+\frac{3\pi}{4}\right)$,其中$t_s=\frac{1}{f_s}=\frac{1}{8000}$, 给定如下$N=8$个采样点，计算其傅里叶变换。

$$x(0)=0.3535, x(1)=0.3535$$

$$x(2)=0.6464, x(3)=1.0607$$

$$x(4)=0.3535, x(5)=-1.0607$$

$$x(6)=-1.3535, x(7)=-0.3535$$



例题1，DFT之后X(m)的幅度

离散傅里叶变换

- DFT的性质

性质1. 对称性，对于实数信号，有

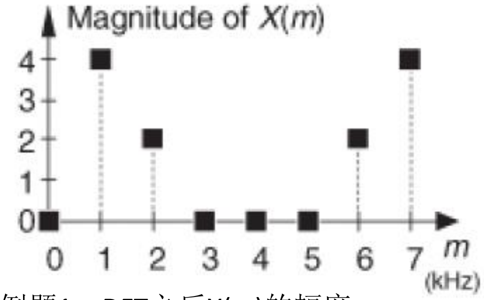
$$X(m) = X^*(N-m)$$

证明：$X(N-m) = \sum_{n=0}^{N-1} x(n)e^{-j2\pi n(N-m)/N}$

$$= \sum_{n=0}^{N-1} x(n)e^{-j2\pi nN/N}\, e^{j2\pi nm/N} = \sum_{n=0}^{N-1} x(n)e^{-j2\pi n} e^{j2\pi nm/N}$$

因为$e^{-j2\pi n} = \cos(2\pi n) - j\sin(2\pi n) = 1$

故$X(N-m) = \sum_{n=0}^{N-1} x(n)e^{j2\pi nm/N} = X^*(m)$



例题1，DFT之后X(m)的幅度

此性质很重要，如上图所示，DFT之后的离散频率序列的幅度具有对称性，因此，在进行N点DFT之后，只需要保留前N/2+1个点。语音信号特征提取时，一般使用512点DFT，由于对称性，我们只需要前257个有效点

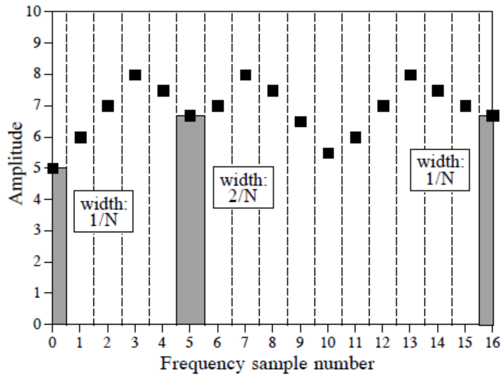
离散傅里叶变换

- DFT的性质

性质2：X(m)实际上表示的是“谱密度”（spectral density），如果对一个幅度为A实正弦波进行N点DFT，则DFT之后，对应频率上的幅度M和A之间的关系为：

$$M = \frac{AN}{2}$$

可以用例题1进行验证！



DFT之后的频域序列X(m)的幅值实际上是一个“密度”的概念，通俗讲，即单位带宽上有多少信号存在。

离散傅里叶变换

- DFT的性质

性质3：DFT的线性

如果$x_{sum}(n) = x_1(n) + x_2(n)$，则对应的频域上有：$X_{sum}(m) = X_1(m) + X_2(m)$

性质4：时移性，对$x(n)$左移k个采样点，得到$x_{shift}(n) = x(n-k)$,对$x_{shift}(n)$进行DFT，有

$$X_{shift}(m) = e^{\frac{j2\pi km}{N}} X(m)$$

离散傅里叶变换

- DFT的频率轴
  - 频率分辨率：f_s/N，表示最小的频率间隔。当N越大时，频率分辨率越高，在频域上，第m个点所表示的分析频率为：

$$f_{analysis}(m) = \frac{m}{N} f_s$$

从这个角度，我们可以理解为$X(m)$的幅值，体现了原信号中频率成分为$\frac{m}{N} f_s$Hz 的信号的强度（性质2）

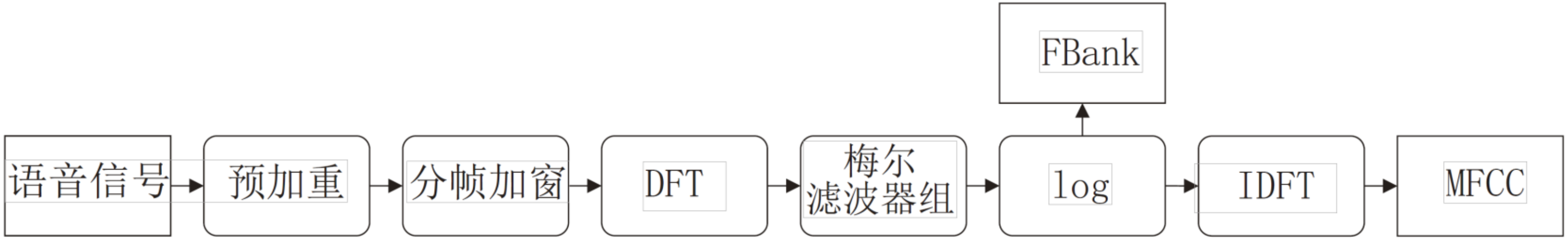
为了提高频率分辨率，我们可以将时域长度为N的信号$x(n)$补0，增加信号的长度，从而提高频率分辨率。对信号进行补0的操作，不会影响DFT的结果，这在FFT（快速傅里叶变换）中和语音信号分析中非常常见。比如，在语音特征提取阶段，对于16k采样率的信号，一帧语音信号长度为400个采样点，为了进行512点的FFT，通常将400个点补0，得到512个采样点，最后只需要前257个点。

离散傅里叶变换

- 快速傅里叶变换（FFT）
  - FFT的基本思想是把原始的N点序列，依次分解成一系列的短序列。充分利用DFT计算式中指数因子 所具有的对称性质和周期性质，进而求出这些短序列相应的DFT并进行适当组合，达到删除重复计算，减少乘法运算和简化结构的目的。
  - 自学FFT算法，推荐教材
    - Understanding DSP，第4章

Fbank和MFCC特征提取

- Fbank和MFCC（Mel-Frequency Cepstral Coefficients）提取流程



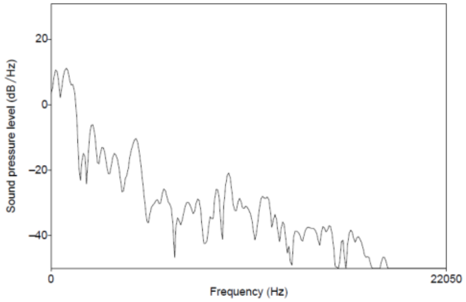
  - Fbank和MFCC特征目前仍是主要使用的特征，虽然有工作尝试直接使用波形建模，但是效果并没有超越基于频域的特征

Fbank和MFCC特征提取

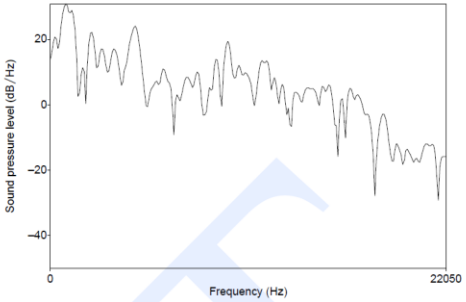
- Step1. 预加重（pre-emphasis）
  - 提高信号高频部分的能量
  - 预加重滤波器是一个一阶高通滤波器，给定时域输入信号$x[n]$，预加重之后的信号为

$$y[n] = x[n] - \alpha x[n-1]$$

其中，$0.9 \leq \alpha \leq 1.0$



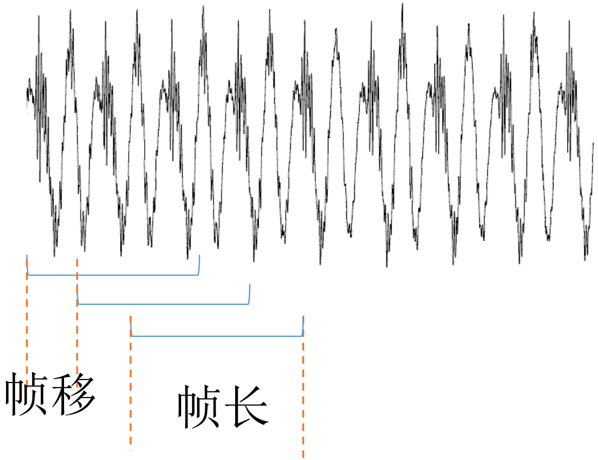
(a)



(b)

Fbank和MFCC特征提取

- Step2. 加窗（windowing）分帧
  - 语音信号为非平稳信号，其统计属性是随着时间变化的
  - 语音信号又具有短时平稳的属性，在进行语音识别的时候，对于一句话，识别的过程也是以较小的发音单元（音素、字音素或者字、字节）为单位进行识别，因此用滑动窗来提取短时片段
  - 帧长、帧移、窗函数，对于采样率为16kHz的信号，帧长、帧移一般为25ms、10ms，即400和160个采样点



Fbank和MFCC特征提取

- Step2. 加窗（windowing）分帧
  - 分帧的过程，在时域上，即是用一个窗函数和原始信号进行相乘

$$y[n] = w[n]x[n]$$

w[n]称为窗函数，常用的窗函数有

矩形窗

$$w[n] = \begin{cases} 1, 0 \le n \le L-1 \\ 0, \quad \text{otherwise} \end{cases}$$

汉明窗 (Hamming)

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L}\right), \ 0 \le n \le L-1 \\ 0, \quad \text{otherwise} \end{cases}$$

Fbank和MFCC特征提取

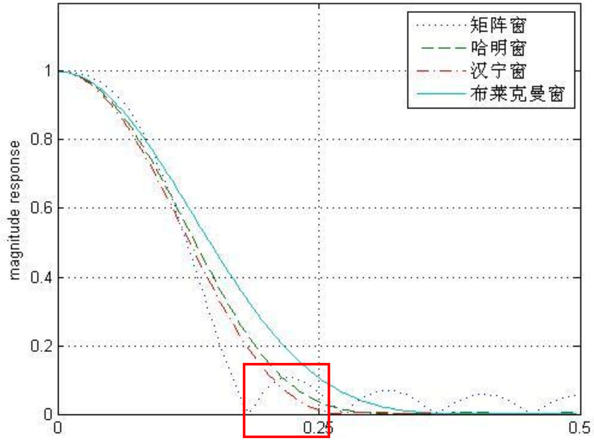
- Step2. 加窗（windowing）分帧
  - 为什么不直接使用矩形窗?

加窗的过程，实际上是在时域上将信号截断，窗函数与信号在时域相乘，就等于对应的频域表示进行卷积（*），矩形窗主瓣窄，但是旁瓣较大（红色部分），将其与原信号的频域表示进行卷积，就会导致频率泄露。

$$y[n] = w[n]x[n]$$

DFT

$$Y=W * X$$

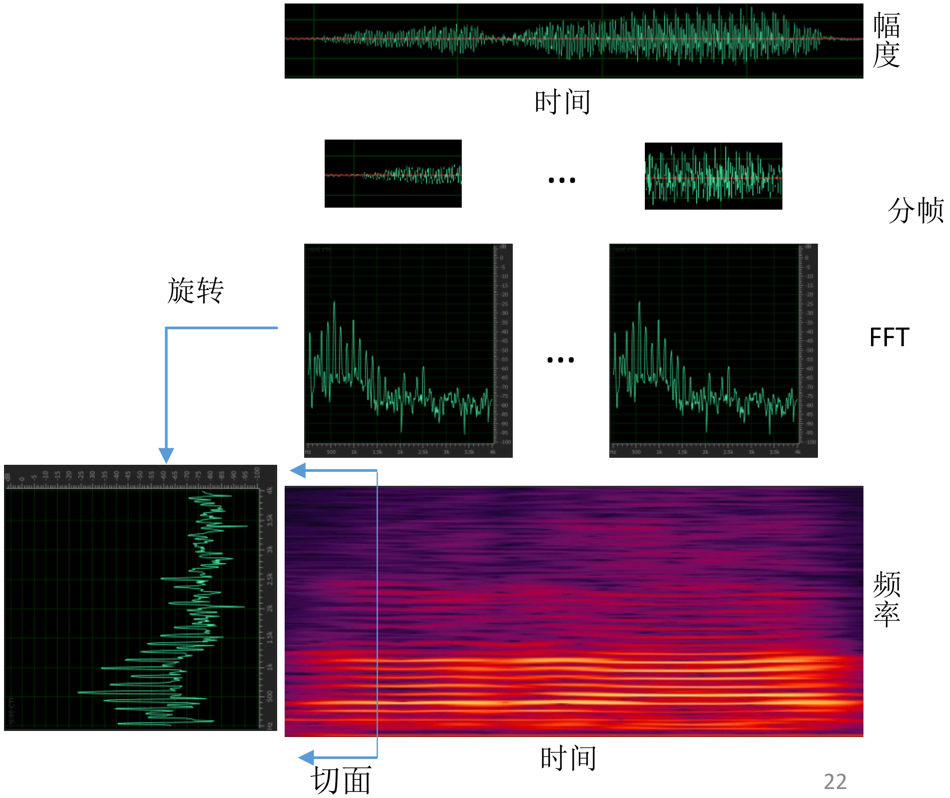


Fbank和MFCC特征提取

- Step3. 傅里叶变换
  - 将上一步分帧之后的语音帧，由时域变换到频域，取DFT系数的模，得到谱特征

右图展示了语谱图的生成过程：

1. 加窗分帧
2. 将每一帧信号进行DFT（FFT），如第t帧信号，DFT系数为$X_t(m), m = 0,1,\dots,N$
3. 将每一帧DFT的系数按时间顺序排列，得到一个矩阵$Y \in C^{T*N}$，且$Y[t,m] = X_t(m)$
4. 语谱图是一个三维图，横轴表示时间(t)，纵轴表示频率，颜色的深浅表示当前时频点上幅度的大小$|Y[t,m]|$

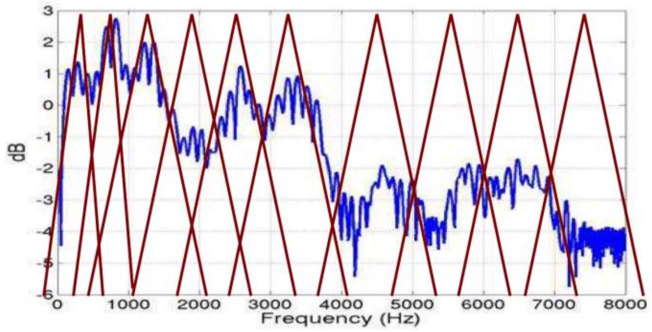


Fbank和MFCC特征提取

- Step4. 梅尔滤波器组和对数操作
  - DFT得到了每个频带上信号的能量，但是人耳对频率的感知不是等间隔的，近似于对数函数
  - 将线性频率转换为梅尔频率，梅尔频率和线性频率转换关系

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

  - 梅尔三角滤波器组：根据起始频率、中间频率和截止频率，确定各滤波器系数

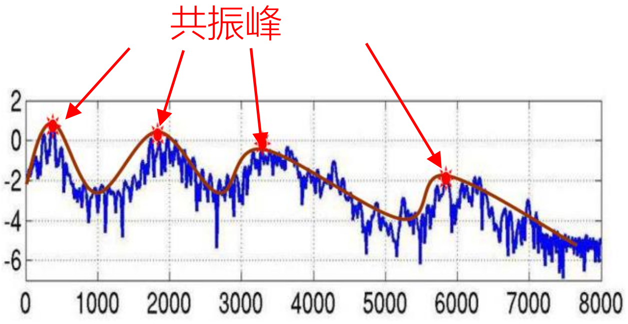


Fbank和MFCC特征提取

- Step4. 梅尔滤波器组和对数操作
  - 梅尔滤波器组设计
    - 确定滤波器组个数P
    - 根据采样率f_s，DFT点数N，滤波器个数P，在梅尔域上等间隔的产生每个滤波器的起始频率、中间频率和截至频率，注意，上一个滤波器的中间频率为下一个滤波器的起始频率（存在overlap）
    - 将梅尔域上每个三角滤波器的起始、中间和截止频率转换线性频率域，并对DFT之后的谱特征进行滤波，得到P个滤波器组能量，进行log 操作，得到Fbank特征
  - MFCC特征在Fbank特征基础上继续进行IDFT变换等操作

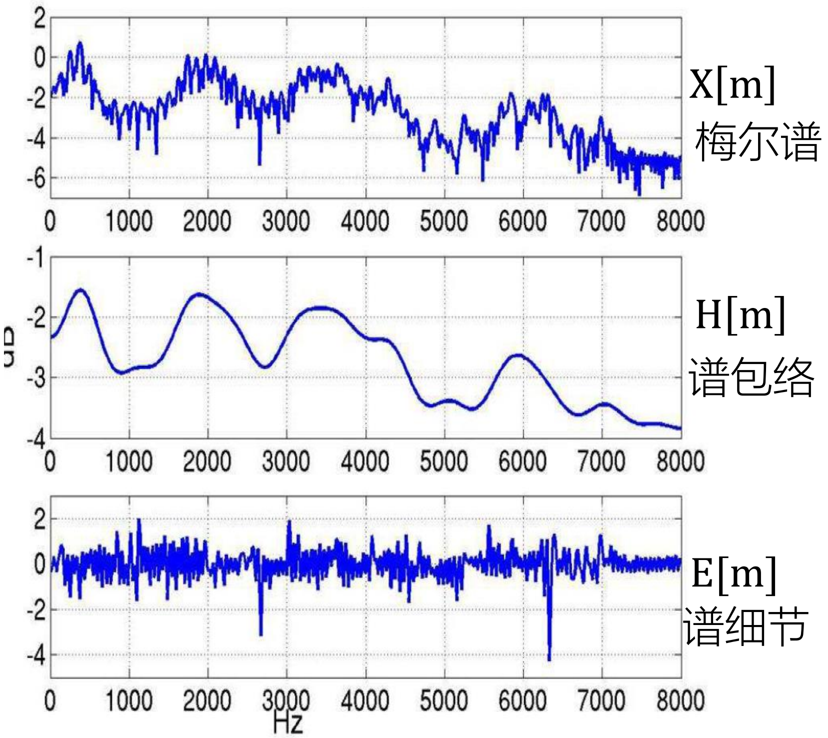
MFCC特征提取

- Step4. 梅尔滤波器组和对数操作
  - 倒谱分析



$$X[m] = H[m]E[m]$$

频域信号可以分解成谱包络(Envelope)和谱细节的乘积，不同音素的谱包络和共振峰具有区分性



MFCC特征提取

- Step4. 梅尔滤波器组和对数操作
  - 倒谱分析

1. $|X[m]| = |H[m]||EM|$

2. $\log|X[m]| = \log|H[m]| + \log|E[m]|$

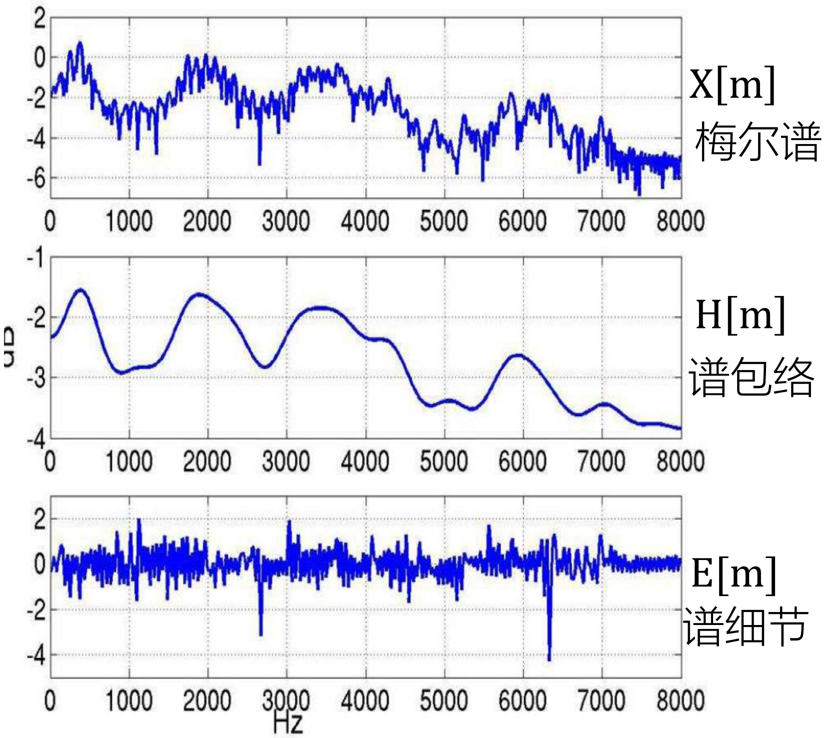
3. 两边进行IDFT （此处为DCT变换）

4. IDFT之后的第1~K个点，为K维MFCC特征

$$c[k] = \sum_{m=0}^{N} (\log|X[m]|)\, e^{j2\pi mk/N}$$

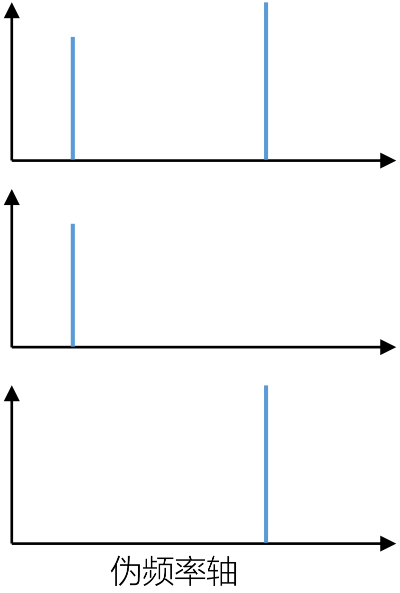
注意：取log有两个目的：

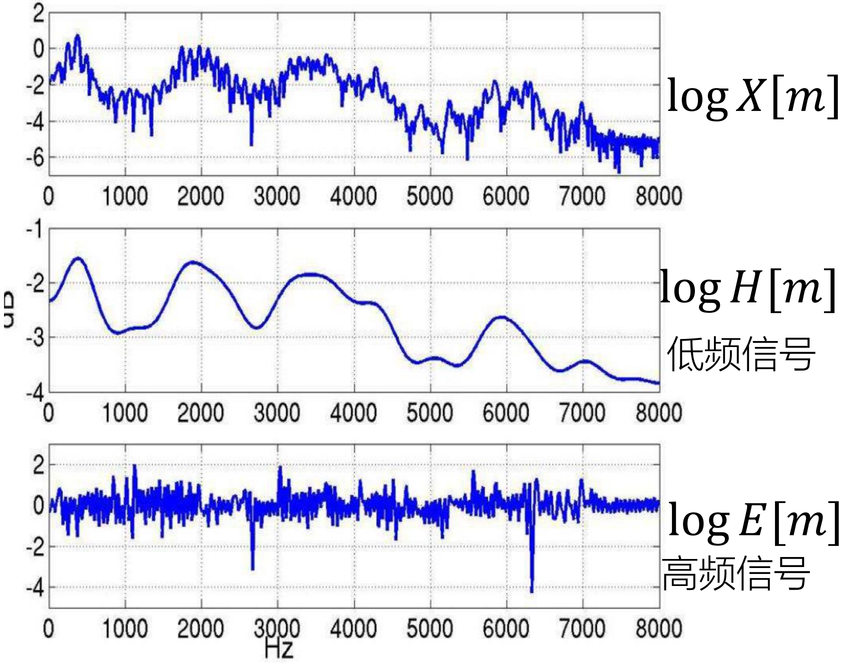
1. 人耳对信号感知是近似对数的，高频部分较为敏感；
2. 对数使特征对输入信号的扰动不敏感



MFCC特征提取

- Step4. 梅尔滤波器组和对数操作
  - IDFT，对数谱的谱！





http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

MFCC特征提取

- Step5. 动态特征计算
  - 一阶差分（Delta, Δ），类比速度，最简单的一阶差分计算方法

$$\Delta(t)=\frac{c(t+1)-c(t-1)}{2}$$

  - 二阶差分（Delta delta, ΔΔ），类比加速度，简单计算方法

$$\Delta\Delta(t)=\frac{\Delta(t+1)-\Delta(t-1)}{2}$$

- Step6. 能量计算

$$e=\sum x^2[n]$$

MFCC特征提取

- MFCC特征总结
  - 一般常用的MFCC特征维39维，包括：
    - 12维原始MFCC
    - 12维Δ
    - 12维$\Delta\ \Delta$
    - 1维能量
    - 1维能量Δ_e
    - 1维能量$\Delta\ \Delta_e$
  - MFCC特征一般用于对角GMM训练，各维度之间相关性小
  - Fbank特征一般用于DNN训练

作业

1. 给定一段音频，请提取12维MFCC特征，阅读代码预加重、分帧、加窗部分，完善作业代码中fbank和mfcc部分，并给出最终的Fbank和MFCC特征，用默认的配置参数，无需进行修改。

https://github.com/nwpuaslp/ASR_Course.git

2. 简答课件第7，9页的问题

3. 提交的压缩包请包含如下文件：

其中test.fbank,test.mfcc为程序生成的输出文件

quiz.txt为对课件问题的回答

- mfcc.py
- quiz.txt
- test.fbank
- test.mfcc
- test.wav

感谢各位聆听！

西工大音频语音与语言处理研究组