

语音特征提取

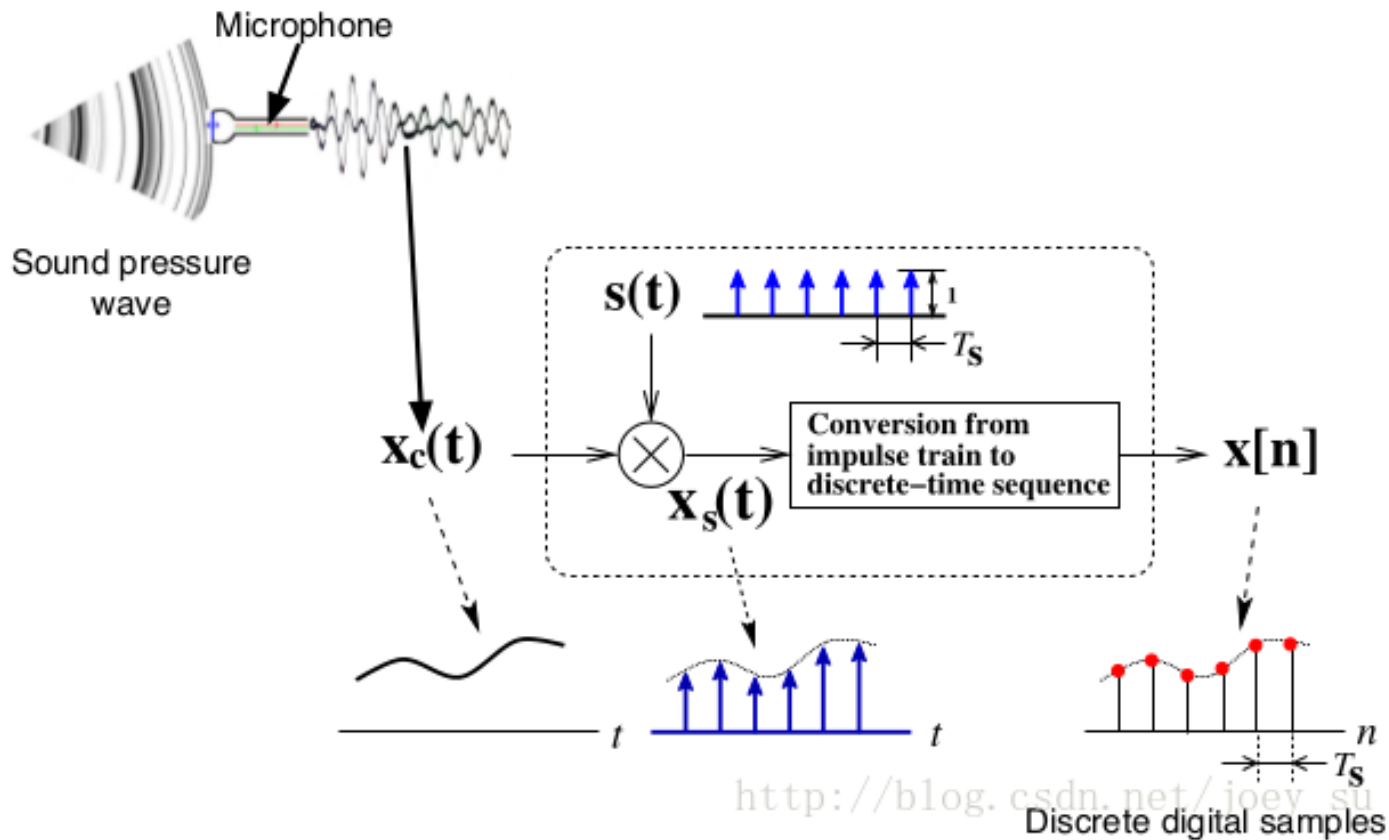
洪青阳 副教授

厦门大学信息科学与技术学院
qyhong@xmu.edu.cn

主要内容

- 1 语音采样与短时分析
- 2 语音信号的频域分析
- 3 人类听觉与Mel频率
- 4 倒谱分析与离散余弦变换
- 5 MFCC特征提取过程
- 6 常用语音声学特征

1 语音采样与短时分析



离散信号

1 语音采样与短时分析

□ 语音信号的短时平稳性

□ 语音信号的短时分析

□ 短时信号的切取

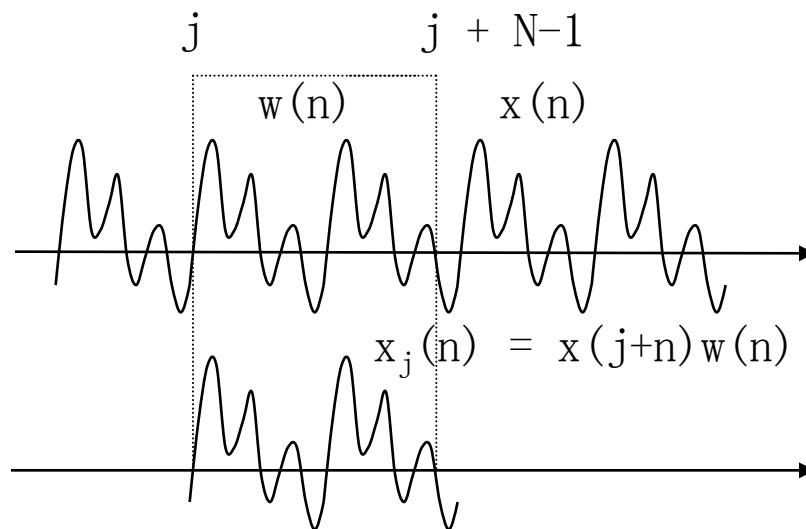
— 分帧

— 帧长

— 帧移

▶ 语音分帧

◦ 每帧10-30ms, 帧间隔10ms



2 语音信号的频域分析





- 语音的感知过程与人类听觉系统具有频谱分析功能紧密相关。因此，对语音信号进行频谱分析，是认识语音信号和处理语音信号的重要方法。
- 声音从频率上可以分为纯音和复合音。纯音只含一种频率的声音(基音)，而没有倍音。复合音是除基音外，还包含多种倍音的声音。**大部分声音(包括语音)都是复合音，涉及多个频率段。**
- 如何分离不同频率信号呢？

傅立叶变换

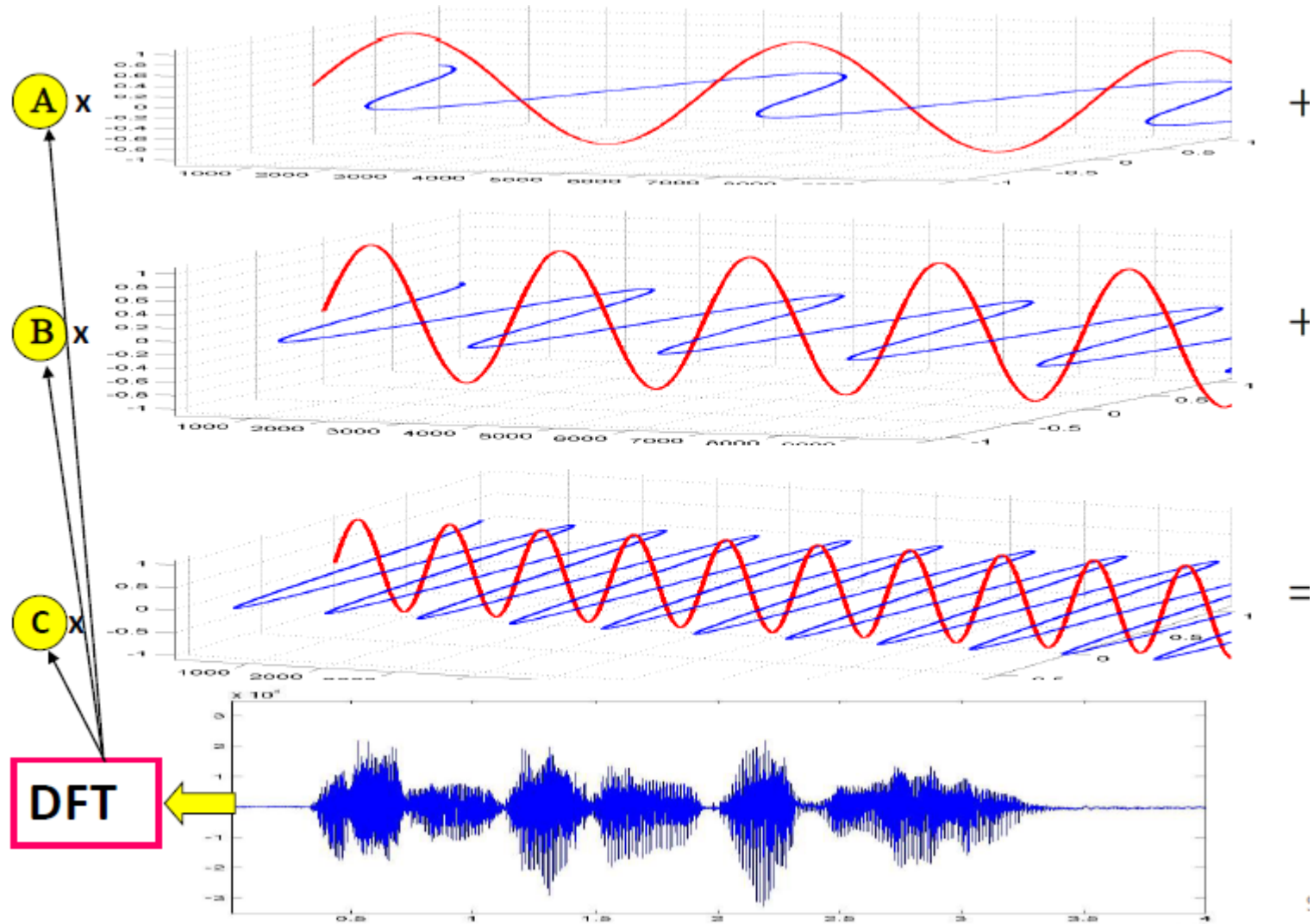
任何连续周期信号都可以由一组适当的正弦曲线组合而成。
---傅立叶(法国数学家)

1	非周期性连续信号
2	周期性连续信号
3	非周期性离散信号
4	周期性离散信号

对有限长序列 $x(i)$ (i :

Type of Transform	Example Signal
Fourier Transform <i>signals that are continuous and aperiodic</i>	
Fourier Series <i>signals that are continuous and periodic</i>	
Discrete Time Fourier Transform <i>signals that are discrete and aperiodic</i>	
Discrete Fourier Transform <i>signals that are discrete and periodic</i>	

离散傅里叶变换(DFT)



Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides

为何用正弦曲线?

- ▶ 正弦波是频域中唯一存在的波形，这是频域中最重要的规则，即正弦波是对频域的描述，因为时域中的任何波形都可用正弦波合成。
- ▶ 一个正弦曲线信号输入后，输出的仍是正弦曲线，只有幅度和相位可能发生变化，但是频率和波的形状仍是一样的。

DFT分解运算方法

- ▶ 利用信号的**相关性(correlation)**可以从噪声背景中检测出已知的信号。
- ▶ 所有的**正弦或余弦**函数是正交的，所以我们可以通过关联的方法把原始信号分离出正余弦信号。

两个不同频率的复指数运算是相互正交的，

$$\int_{-\infty}^{\infty} e^{j\alpha t} e^{-j\beta t} dt = 0, \quad \text{if } \alpha \neq \beta$$

复数形式傅立叶变换

- ▶ 复数形式傅立叶变换把原始信号 $x[n]$ 当成是一个用复数来表示的信号，其中实数部分表示原始信号值，虚数部分为0。因此需要乘以一个复数形式的正交函数。

离散傅里叶变换(DFT)

角频率 $\omega = 2\pi k/K$, 取值范围是 $0 \sim 2\pi$;

傅里叶变换的第 k 个点计算如下:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi kn}{K}} = \sum_{n=0}^{N-1} x[n] e^{-j\omega n}$$

其中

$x[n]$ 是时域波形第 n 个采样点, $X[k]$ 是傅里叶频谱第 k 个点, N 是采样系列里的点数, K 是DFT的大小, 其中 $K \geq N$ 。

DFT系数通常是复数的, 即

欧拉等式: $e^{j\omega n} = \cos(\omega n) + j\sin(\omega n)$

$$e^{-j\frac{2\pi kn}{K}} = \cos\left(\frac{2\pi kn}{K}\right) - j\sin\left(\frac{2\pi kn}{K}\right)$$

变换结果 $X[k]$ 也是个复数的形式, 但这里的虚数部分是有值的。

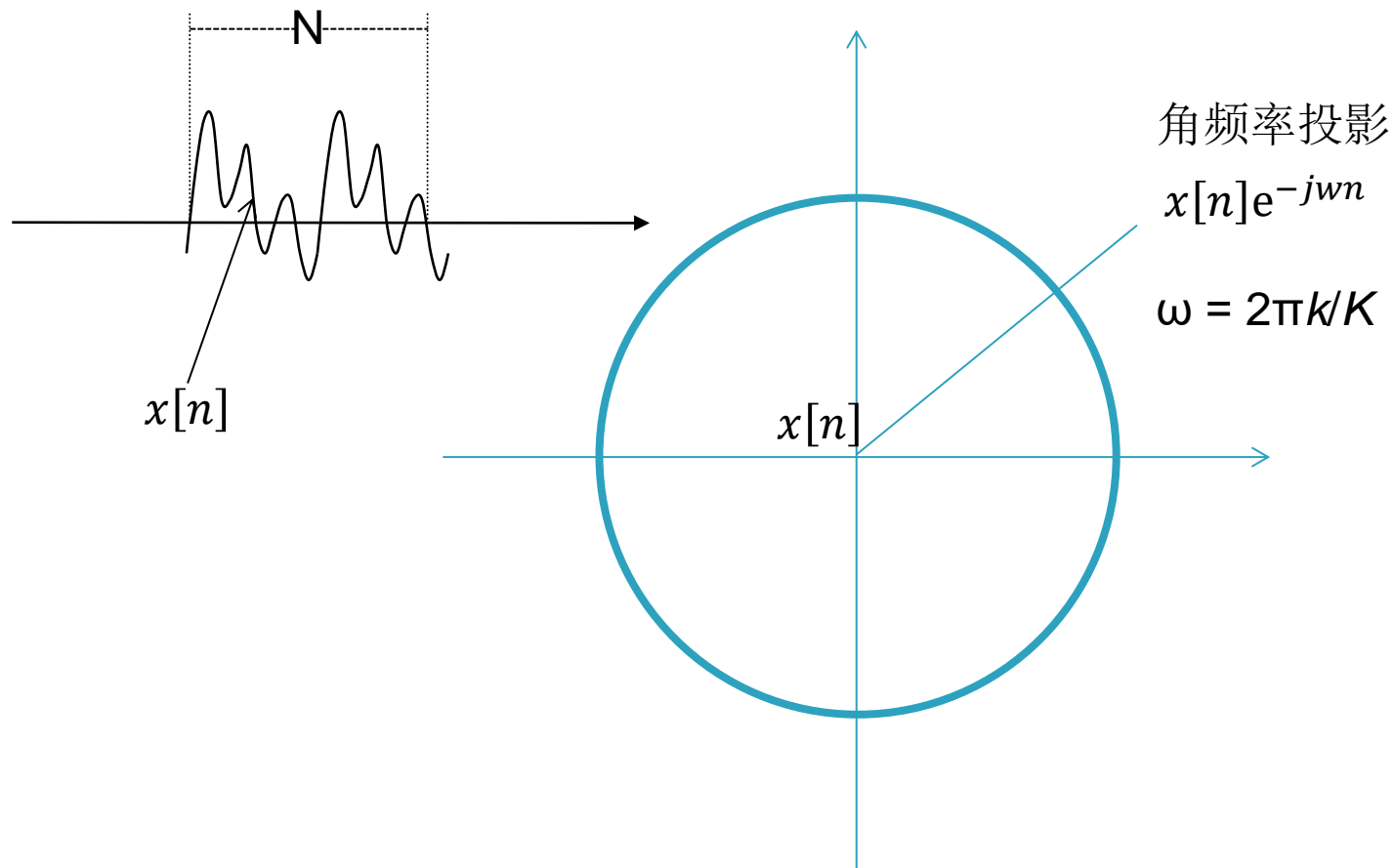
则

$$X[k] = X_{\text{real}}[k] - jX_{\text{imag}}[k]$$

其中

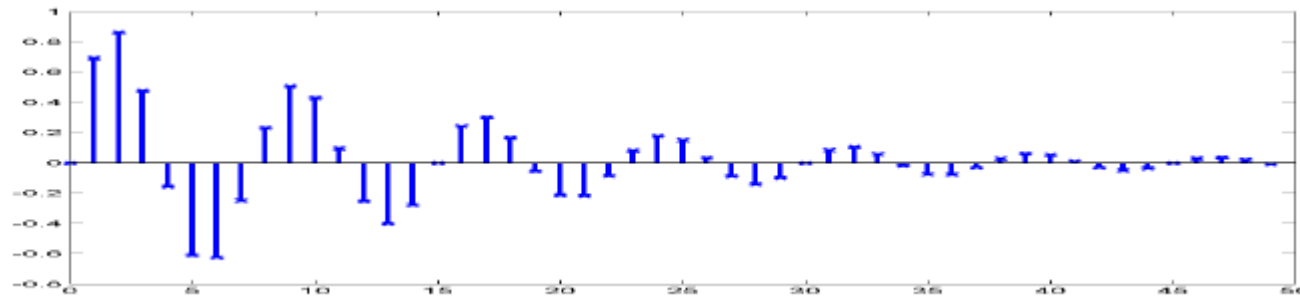
$$X_{\text{real}}[k] = \sum_{n=0}^{N-1} x[n] \cos\left(\frac{2\pi kn}{K}\right)$$
$$X_{\text{imag}}[k] = \sum_{n=0}^{N-1} x[n] \sin\left(\frac{2\pi kn}{K}\right)$$

离散傅里叶变换(DFT)



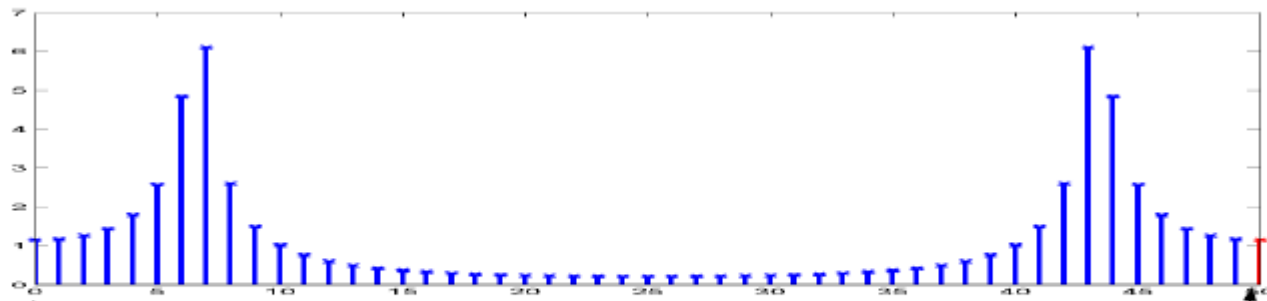
第k点频谱 $X[k] = \sum_{n=0}^{N-1} x[n]e^{-j\omega n}$

离散傅里叶变换(DFT)



N=50

N=400



K=50

K=512

Sample 0 = 0 Hz

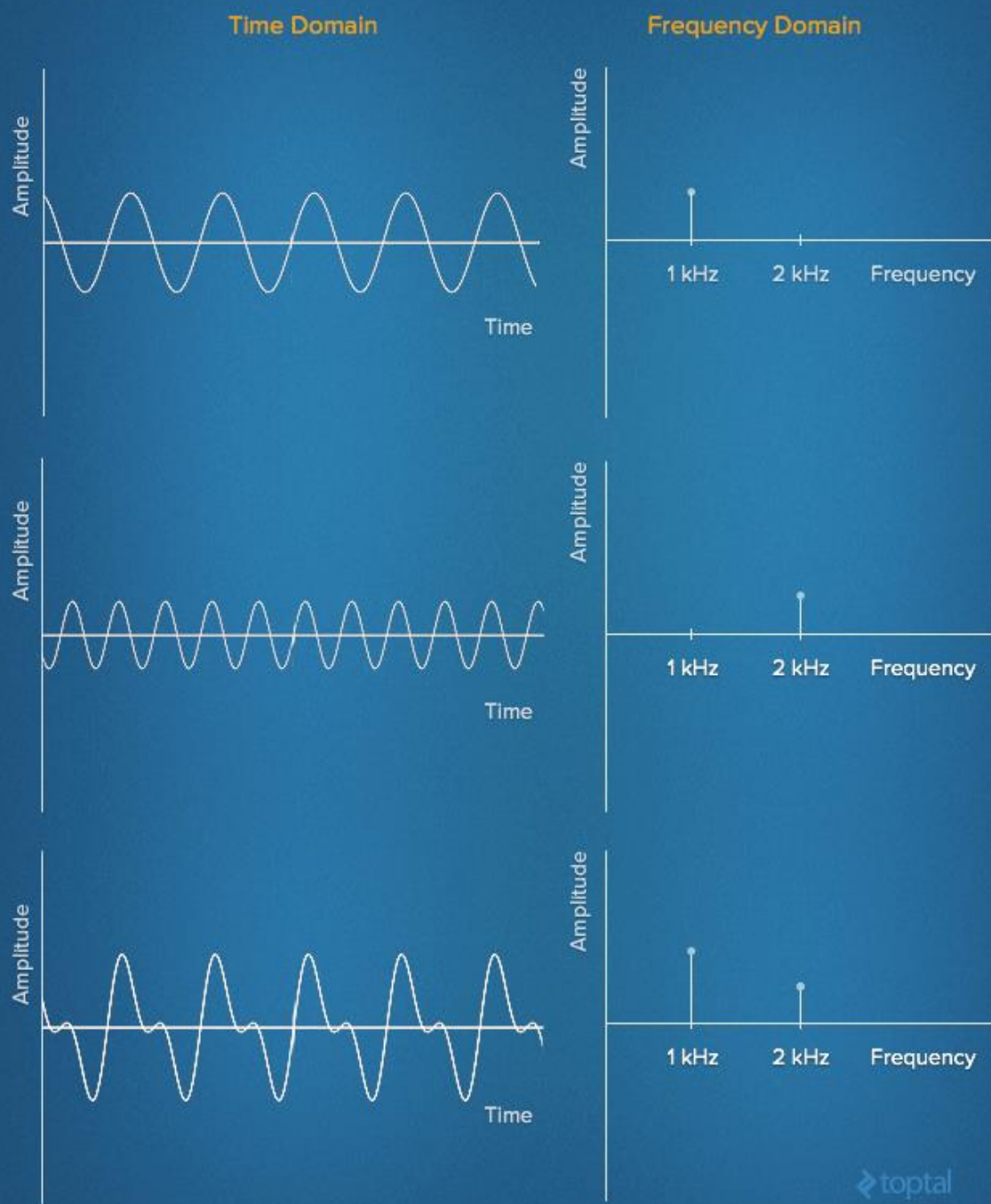
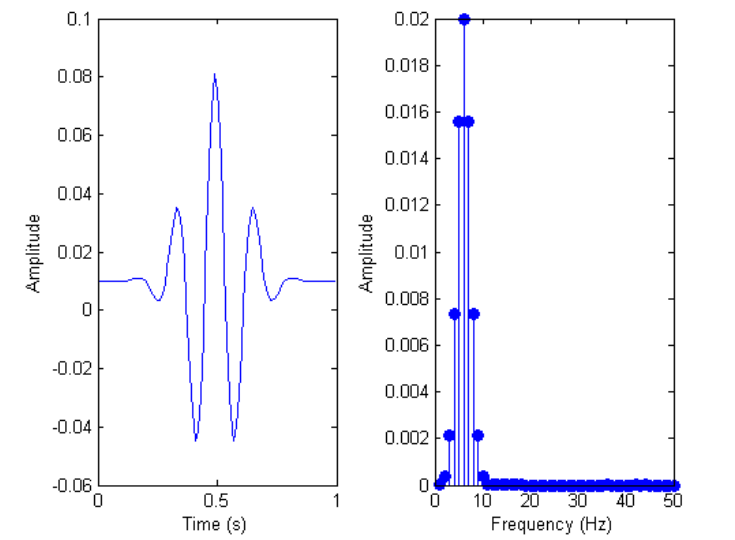
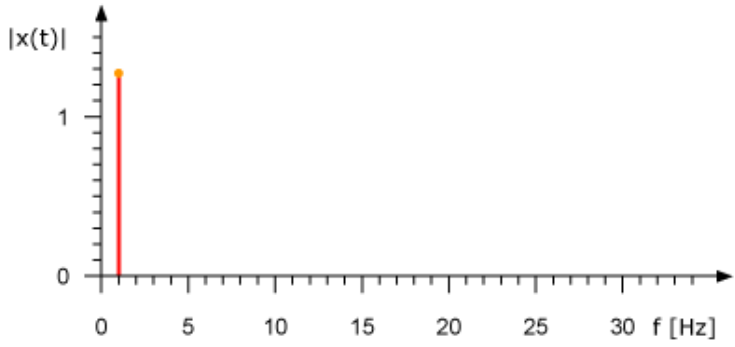
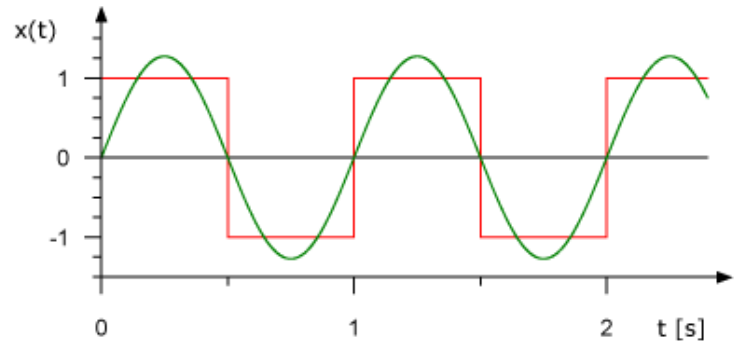
Sample 50 is the 51st point
It is identical to Sample 0

Sample 50 = 8000Hz

Sample 512 = 16000Hz

- ◆ 时域波形的 N 个点经DFT后, 对应 K 个点。
- ◆ 第0个点代表0Hz, 第 $K - 1$ 个点代表 $(K - 1)/K * \text{采样率} F_s$ 。 K 个点在频率轴平均分布。

Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides



离散傅里叶变换(DFT)

振幅频谱用傅里叶系数的幅度表示如下：

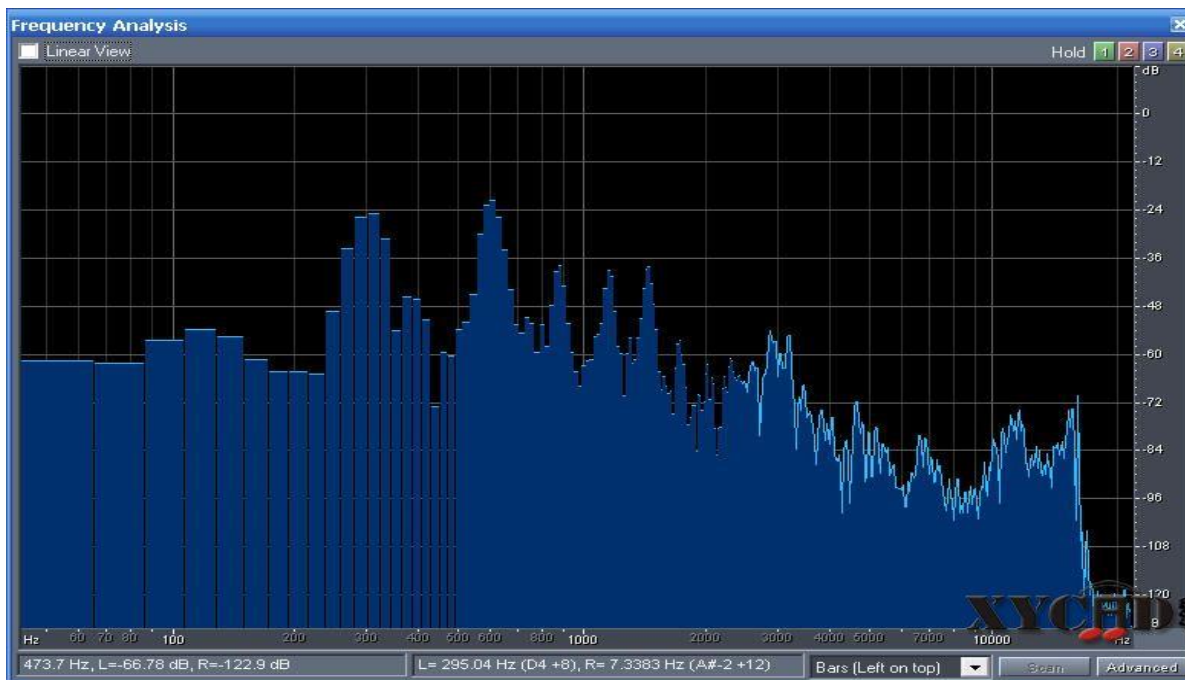
$$\begin{aligned} X_{\text{magnitude}}[k] \\ &= \text{sqrt}(X_{\text{real}}[k]^2 + X_{\text{imag}}[k]^2) \end{aligned}$$

能量频谱用振幅频谱的平方：

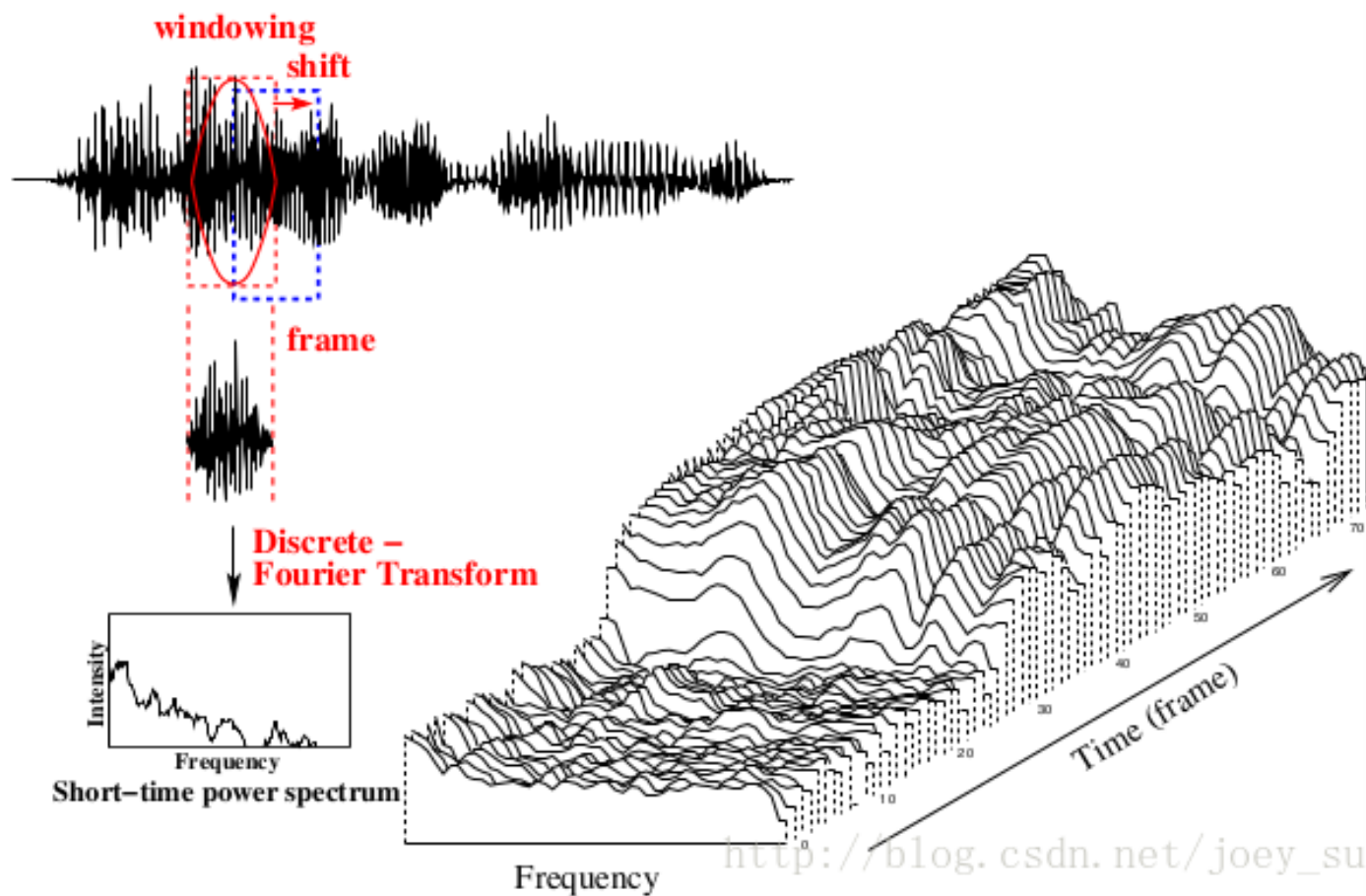
$$X_{\text{power}}[k] = X_{\text{real}}[k]^2 + X_{\text{imag}}[k]^2$$

频谱图(二维)

- ▶ 各种声源发出的声音大多是由许多不同强度、不同频率组成的复合音。在复合音中，不同频率成分的声波具有不同的能量，这种频率成分与能量分布的关系称为声音的**频谱(frequency spectrum)**。各频率成分与能量分布关系的图形称为**频谱图**。

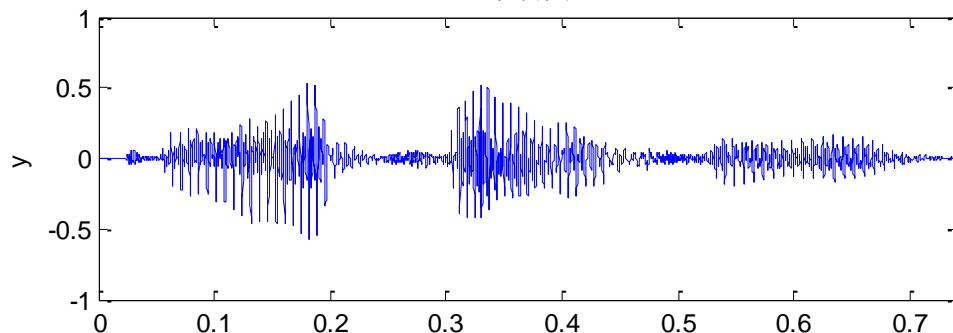


离散傅里叶变换(DFT)

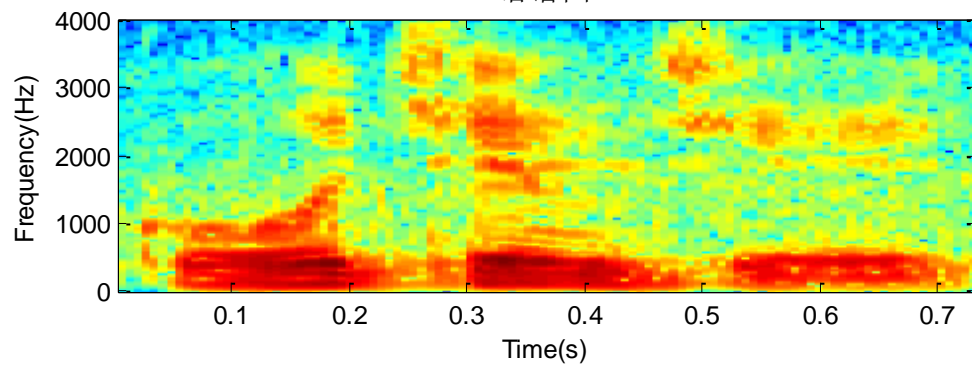


语谱图(三维)

时域图

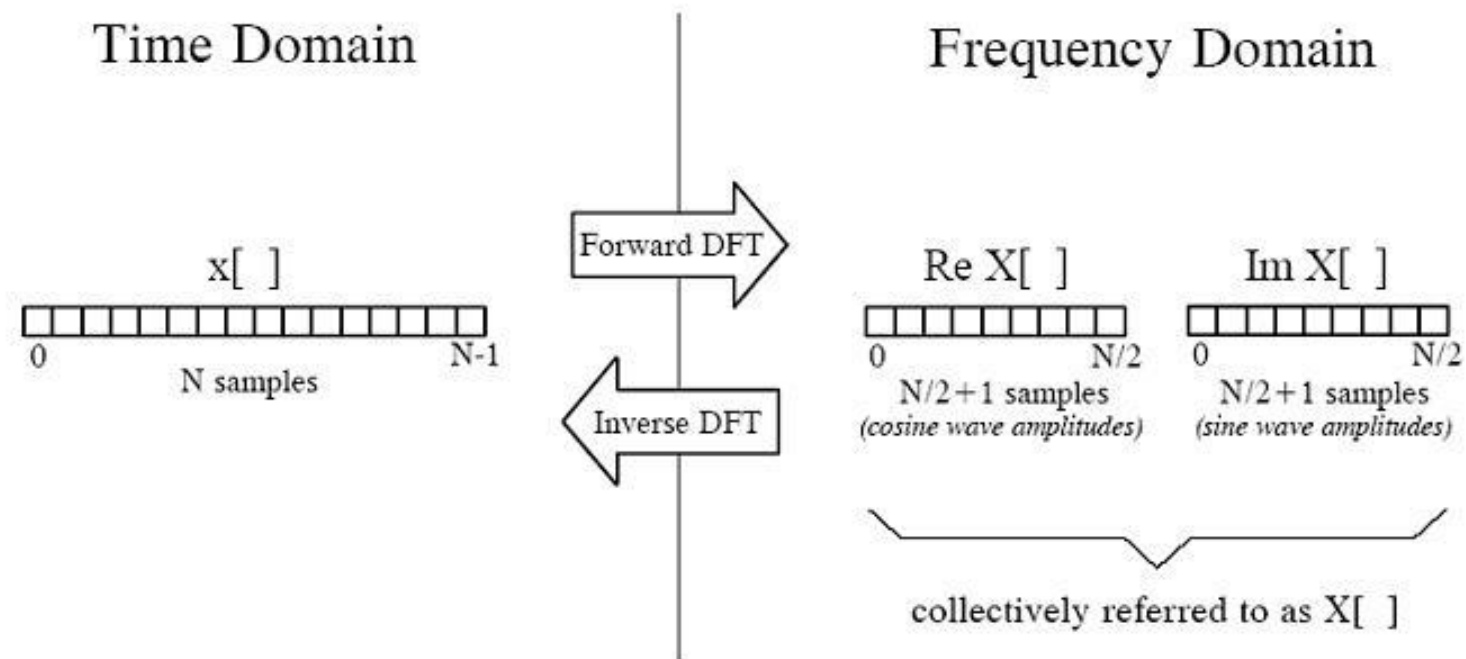


语谱图



语谱图

离散傅里叶变换(DFT)



时域信号可以从DFT恢复得到：

$$x[n] = \frac{1}{K} \sum_{k=0}^{K-1} X[k] e^{\frac{j2\pi kn}{N}}$$

快速傅里叶变换(FFT)

- ◆ 快速傅里叶变换（FFT）是DFT的快速算法。
- ◆ 它是根据DFT的奇、偶、虚、实等特性，对DFT的算法进行改进获得的。
- ◆ FFT对傅立叶变换的理论并没有新的发现，但是对于在计算机系统或者说数字系统中应用离散傅立叶变换，可以说是进了一大步。

3 人类听觉与Mel频率

生理	感知	物理解释
强度	响度	音量是声音的强弱，跟响度不成正比。
基本频率(F0)	音高	指声音的高低，即音调。音高的变化——声调（阴平、阳平、上声、去声）
频率	音调	音调和频率并不是成正比的关系，它还与声音的强度及波形有关
频谱形状	音色	共振频率(F1、F2、F3)
开合时间	音长	声音的长短
双耳相位差	声源位置	

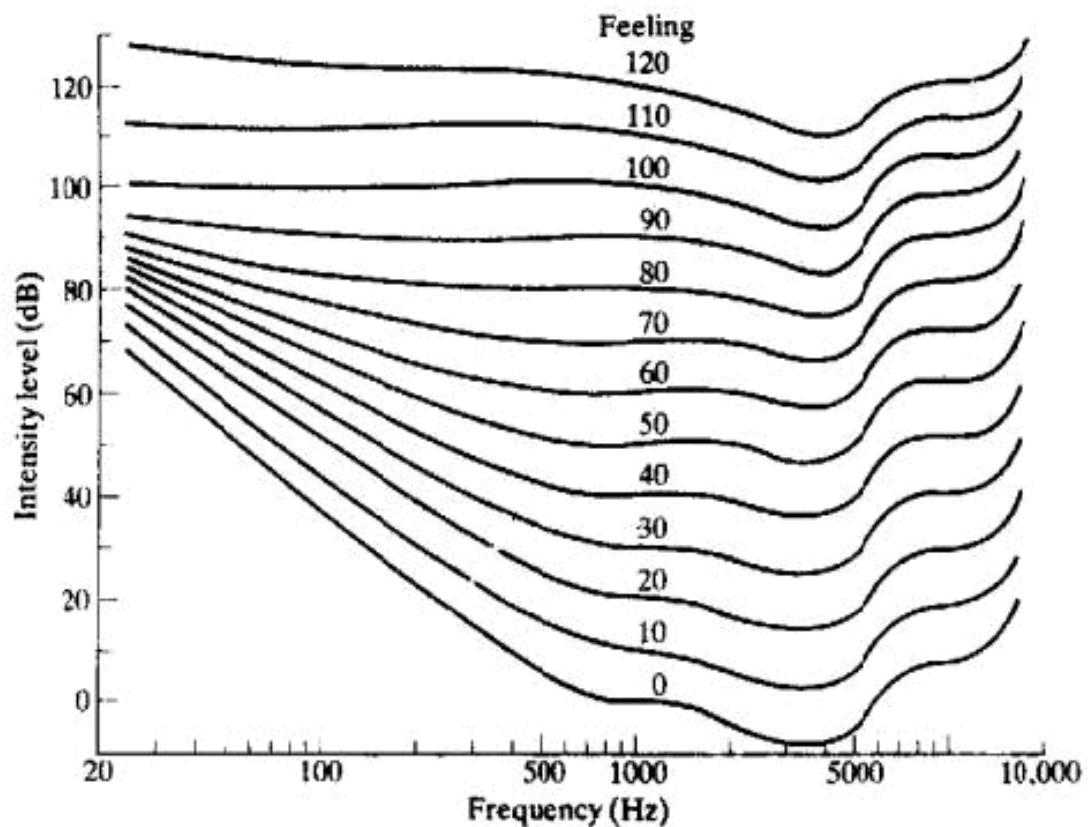
技术术语：

- 等响度轮廓
- Mel频率

音调

- ▶ 音调是听觉分辨声音高低时，用于描述这种感觉的一种特性。客观上用频率来表示音调，主观上感觉音调的单位用**Mel频率**。
- ▶ 一般对于频率低的声音，听起来觉得它的音调低，而频率高的声音，听起来感觉它的音调高。但是**音调和频率并不是成正比的关系，它还与声音的强度及波形有关。**

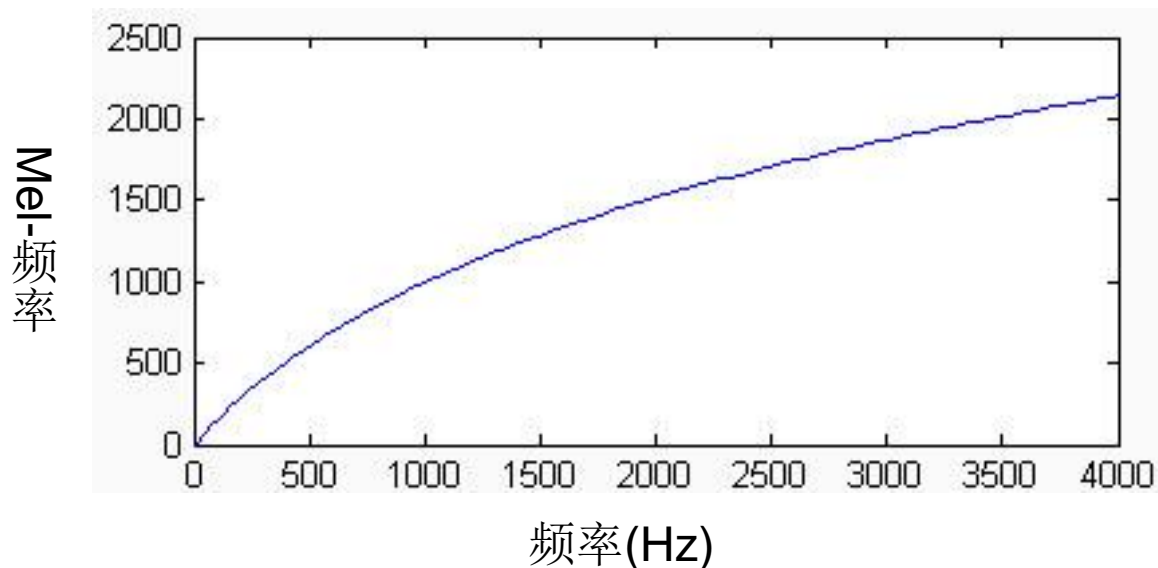
等响度轮廓



等响度曲线与声强级的关系

Mel频率

- ▶ 目的：模拟人耳对不同频率语音的感知
- ▶ 人类对不同频率语音有不同的感知能力
 - 1kHz以下，与频率成线性关系
 - 1kHz以上，与频率成对数关系



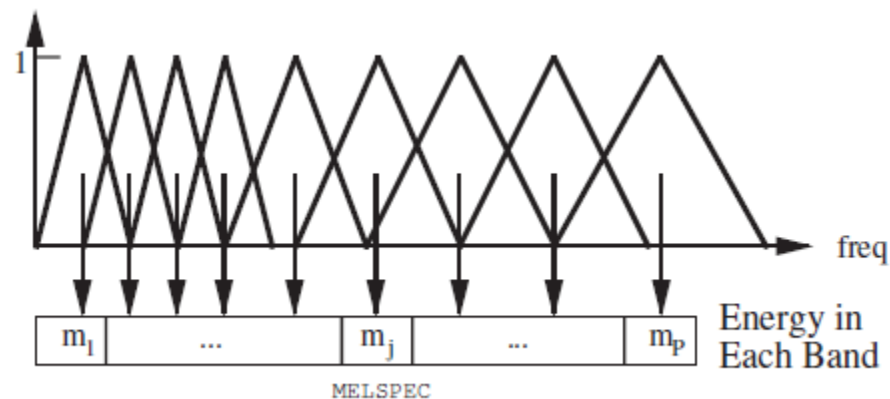
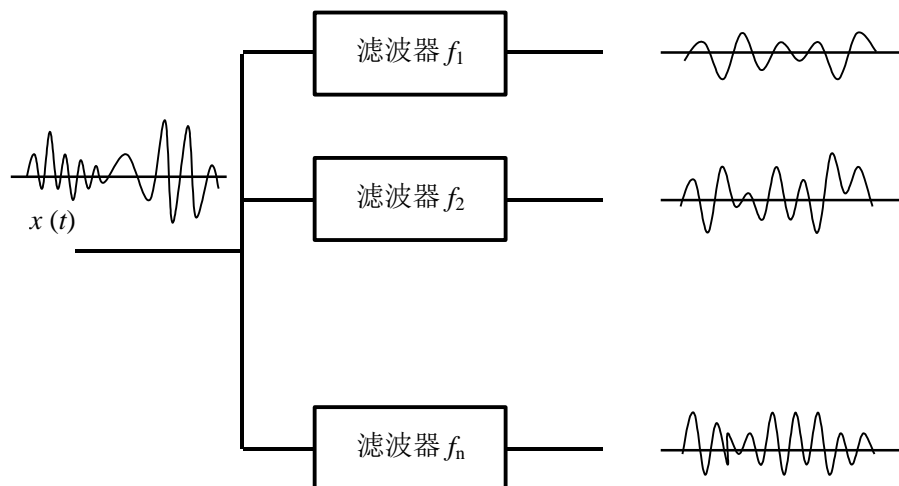
Mel频率

- ▶ Mel频率定义
 - 1 Mel—1 kHz音调感知程度的1 / 1000
- ▶ Mel频率可以用公式表达如下：

$$\begin{aligned}\text{Mel}(f) &= 2595 * \log_{10}(1+f/700) \\ &= 1127 * \ln(1+f/700)\end{aligned}$$

Mel频率滤波器组

- ▶ 根据人耳对低频信号比对高频信号更敏感这一原则，研究者根据心理学实验得到了类似于耳蜗作用的一组滤波器组，这就是Mel频率滤波器组。
- ▶ 小于1000Hz的部分为**线性间隔**，而大于1000Hz的部分为**对数间隔**。

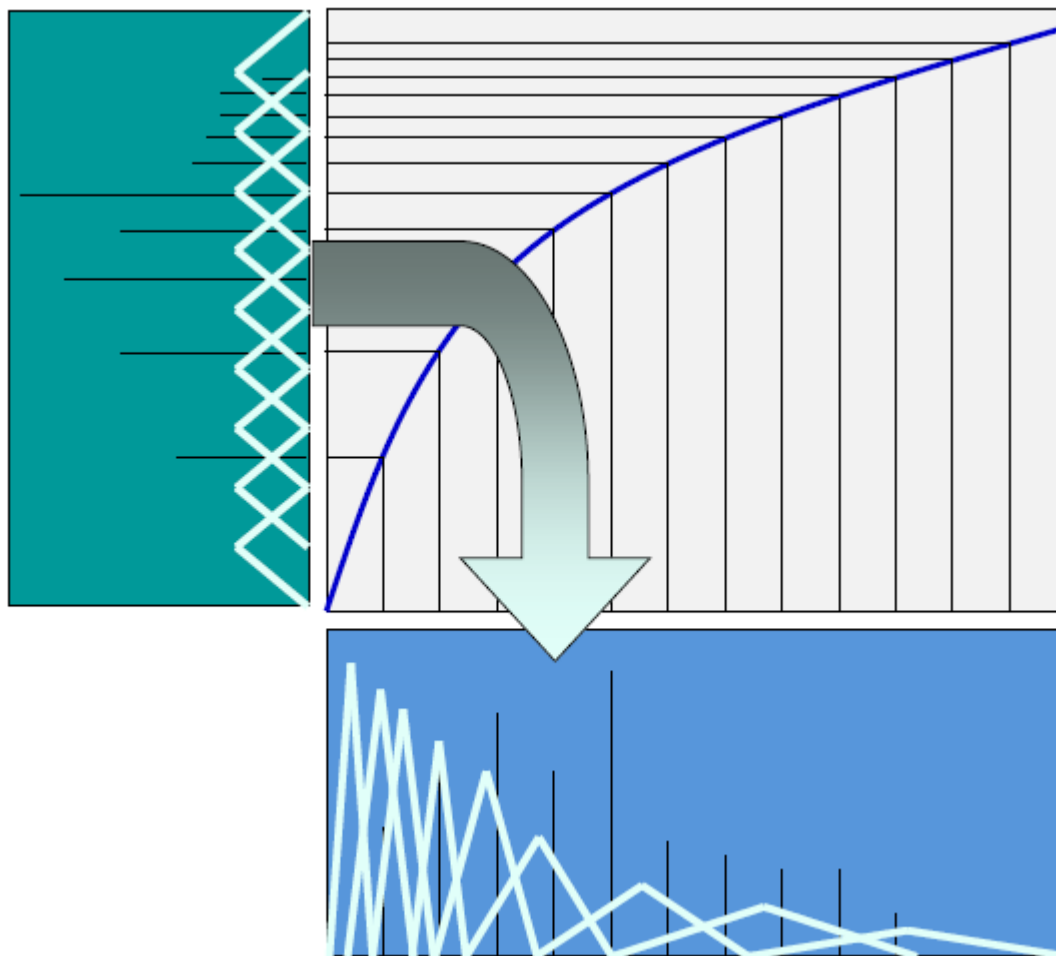


短时帧长：400

补零

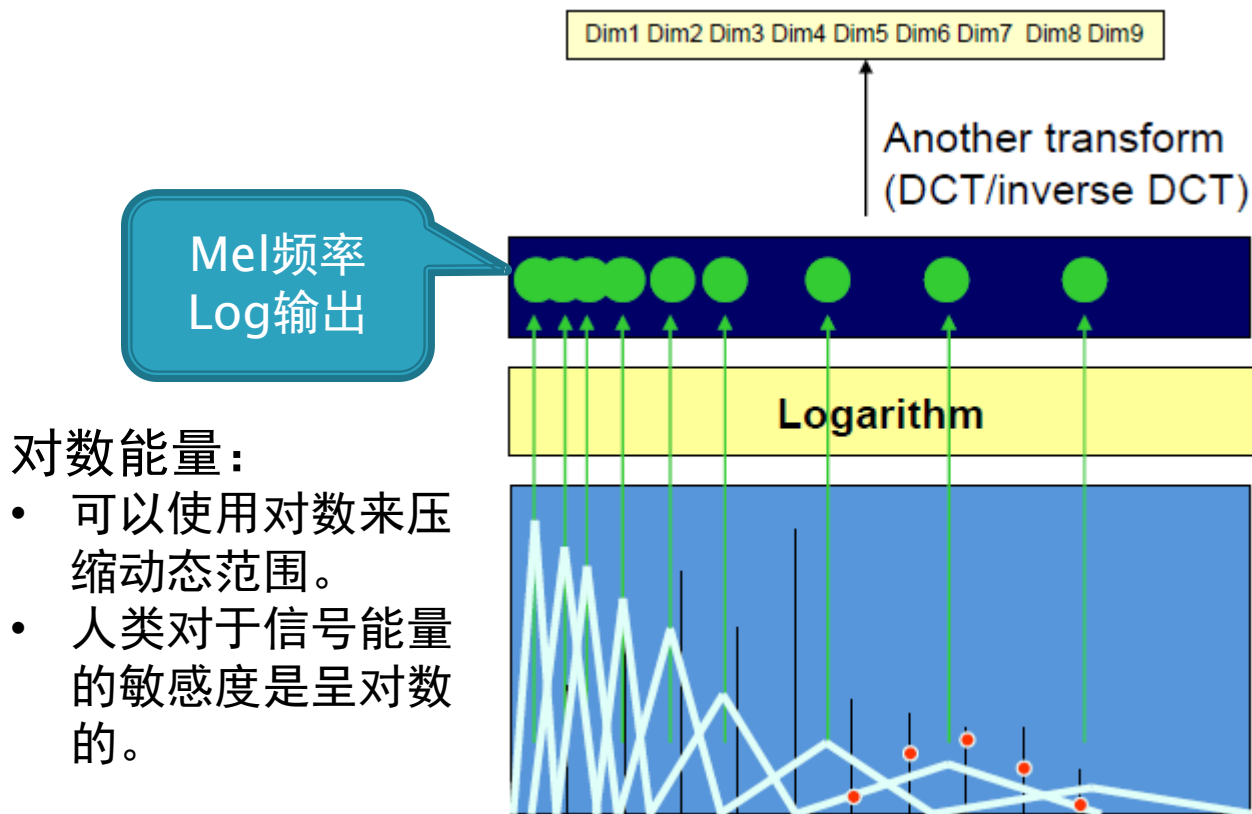
FFT帧长：512

Mel频率滤波器组



Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides

将每个滤波器的输出取对数



对数能量：

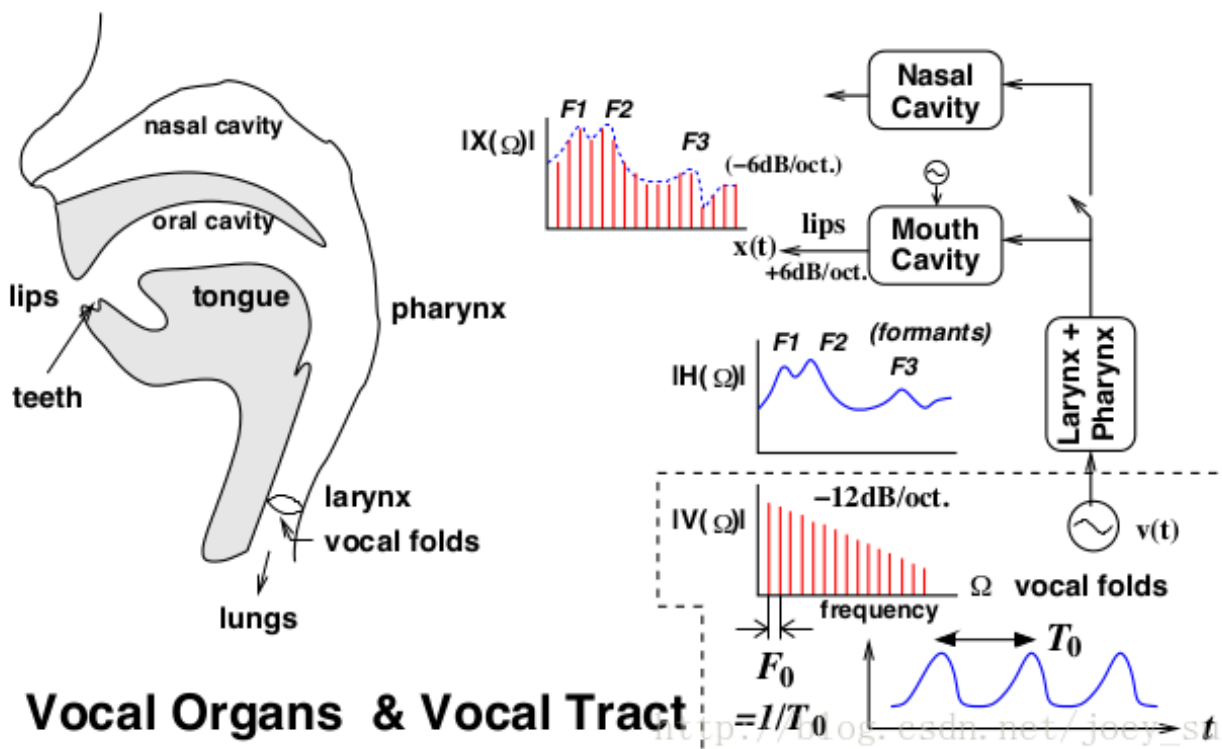
- 可以使用对数来压缩动态范围。
- 人类对于信号能量的敏感度是呈对数的。

Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides

4 倒谱分析与离散余弦变换(DCT)

▶ 语音产生模型:

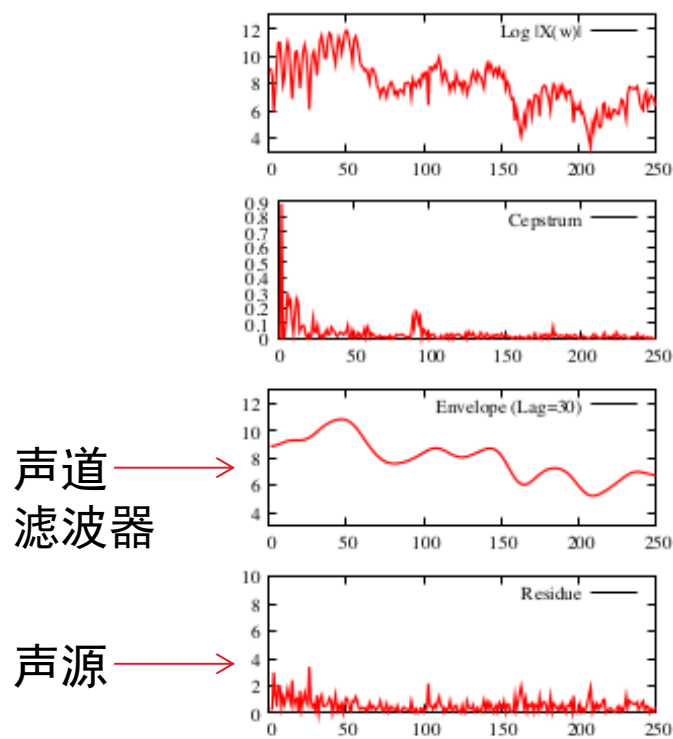
- 发声源 (Source) : 声带振动产生声门源波形
- 滤波器 (Filter) : 发声源波形通过声道(舌头的位置, 下巴等)。



倒谱(Cepstrum)分析

▶ 分离发声源和滤波器

- 发声源：声源基频 F_0 ，动态的声门脉冲
- 滤波器：声道特征，区分音素(phone)



Log Spectrum (freq domain)

↓ Inverse Fourier Transform

Cepstrum (time domain) (quefrequency)

↓ Liftering to get low/high part
(lifter: filter used in cepstral domain)

↓ Fourier Transform

Smoothed-spectrum (freq. domain)
[low-part of cepstrum]

Log spectrum
[high-part of cepstrum]

http://blog.csdn.net/joey_su

倒谱是通过对对数幅度谱进行逆离散傅立叶变换而得到的，即**离散余弦变换(DCT)**

离散余弦变换(DCT)

- ▶ 离散余弦变换(Discrete Cosine Transform, DCT)是与傅里叶变换相关的一种变换，类似于离散傅里叶变换，但是只使用了实数。
- ▶ 离散余弦变换相当于一个长度大概是它两倍的离散傅里叶变换，这个离散傅里叶变换是对一个实偶函数进行的（因为一个实偶函数的傅里叶变换仍然是一个是偶函数）。

4 离散余弦变换

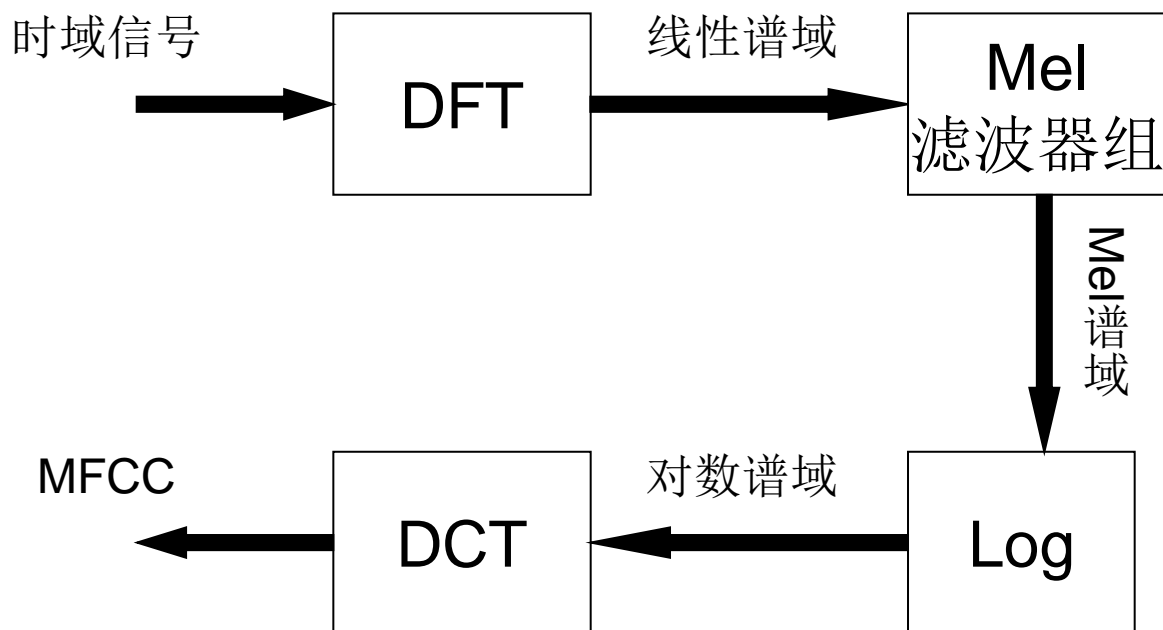
离散余弦变换是根据下面的公式把M个滤波器输出的实数 $x(0), x(1), \dots, x(M-1)$ 变换到另外L个实数 C_0, C_1, \dots, C_{L-1} 的操作

$$C_n = \sum_{k=1}^M \log x'(k) \cos[\pi(k - 0.5)n / M] \quad n = 1, 2, \dots, L$$

4 离散余弦变换

- ▶ 由于离散余弦变换具有很强的“能量集中”特性：大多数的自然信号(包括声音和图像)的能量都集中在离散余弦变换后的低频部分。
- ▶ 离散余弦变换具有最优的去相关性的性能，所以在信号处理中得到广泛应用。

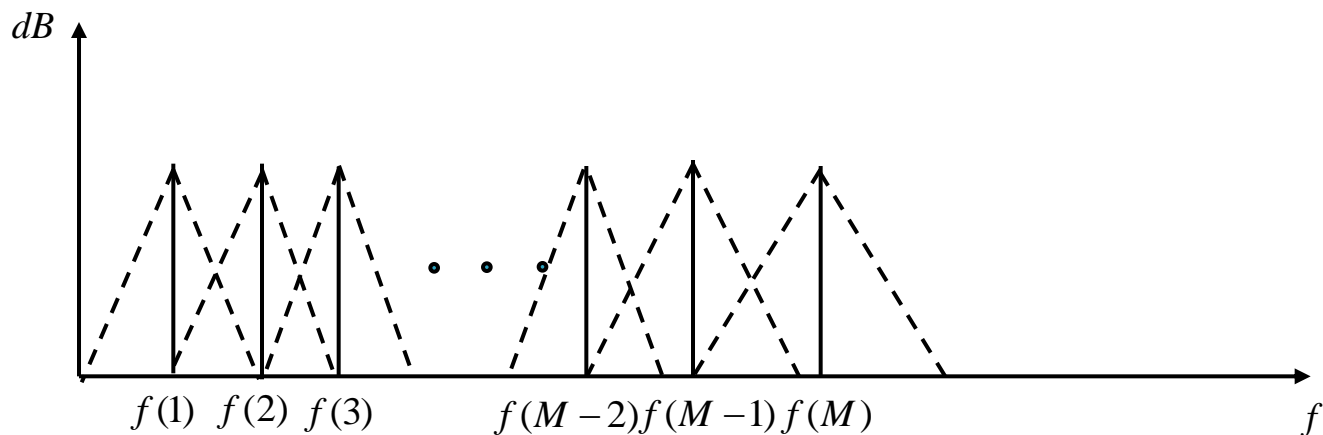
5. Mel频率倒谱系数 (Mel Frequency Cepstrum Coefficient, MFCC)



MFCC计算过程

(1) 将信号进行分帧，预加重和加汉明窗处理，然后进行短时傅立叶变换得到其频谱；

(2) 求频谱平方，即能量谱，将每个滤波频带内的能量进行叠加，第 k 个滤波器输出功率谱 $x'(k)$

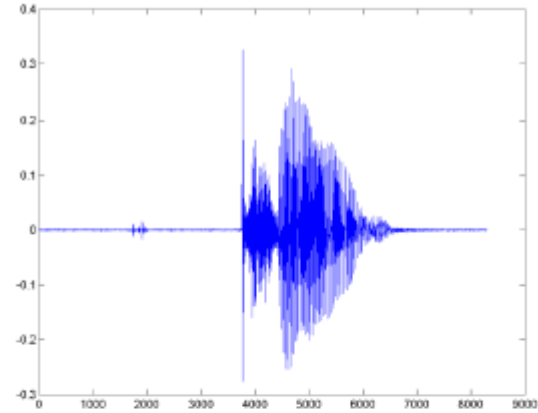
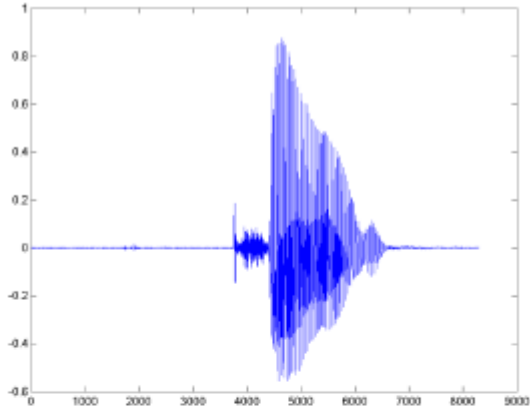


(3) 将每个滤波器的输出取对数，得到相应频带的对数功率谱；并进行反离散余弦变换，得到 L 个MFCC系数，

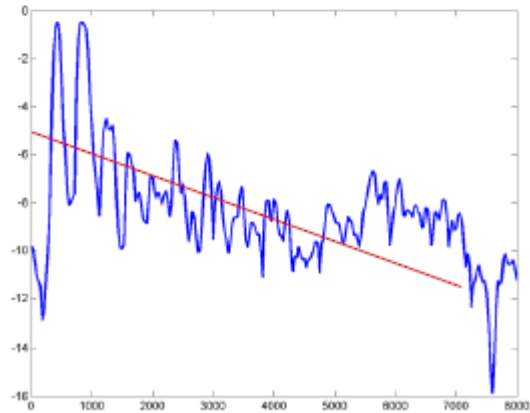
$$C_n = \sum_{k=1}^M \log x'(k) \cos[\pi(k - 0.5)n / M] \quad n = 1, 2, \dots, L$$

(4) 这种直接得到的MFCC特征作为静态特征，将这种静态特征做一阶和二阶差分，得到相应的动态特征。

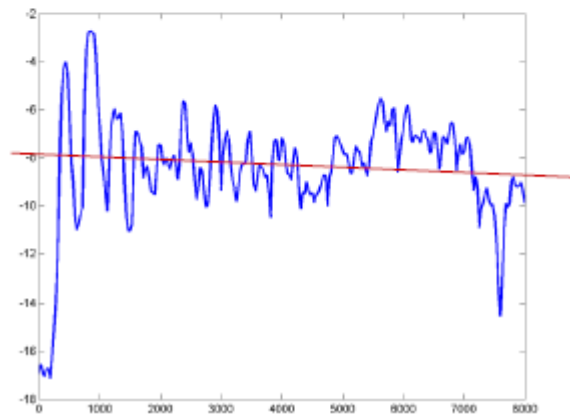
预加重(Preemphasize)



Log(average(magnitude spectrum))



Log(average(magnitude spectrum))



Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides

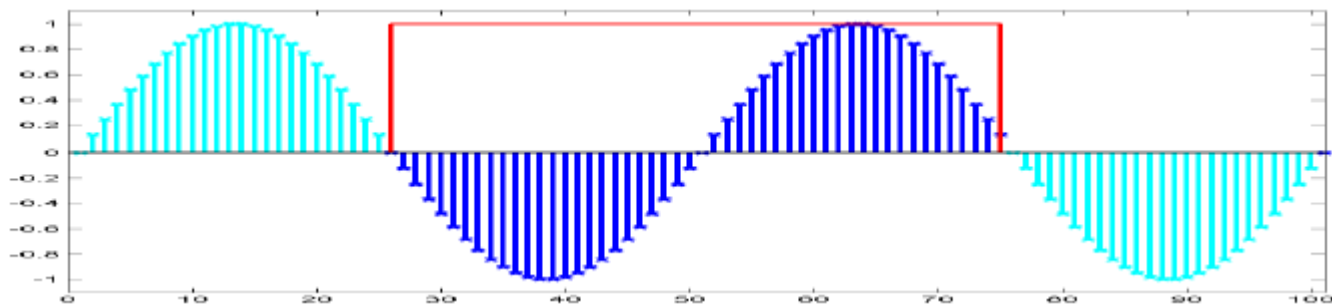
预加重(Preemphasize)

- ▶ 语音是由声门激励通过系统（声道等）产生的，声门激励属于低频，所以语音的能量主要集中在低频，相对于低频来说，高频的能量较低。
- ▶ 预加重的目的就是提升语音信号高频部分，使信号频谱变得比较平坦，保持在低频到高频的整个频带中，能用同样的信噪比求频谱，便于进行频谱分析或者声道参数分析。

$$x'[n] = x[n] - \alpha x[n - 1], 0.95 < \alpha < 0.99$$

加窗(Windowing)

- ▶ 为了得到短时的音频信号，要对音频信号进行加窗操作。窗函数平滑地在音频信号上滑动，将音频信号分成帧。
- ▶ 在加窗的时候，不同的窗口选择将影响到音频信号分析的结果。在选择窗函数时，一般有两个问题要考虑。第一个问题是窗口的形状，即窗函数的形式。第二个问题是窗口的长度。



典型的窗函数

▶ 矩形窗：
$$W_R = \begin{cases} 1 & (0 \leq n < N-1) \\ 0 & (\text{Other}) \end{cases}$$

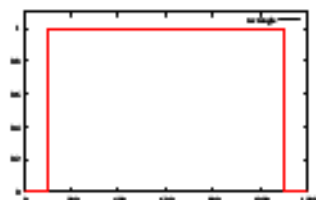
▶ 汉明窗 (Hamming) :

$$W_{HM} = \begin{cases} 0.54 - 0.46 \cos(2\pi n / (N-1)) & (0 \leq n < N-1) \\ 0 & (\text{Other}) \end{cases}$$

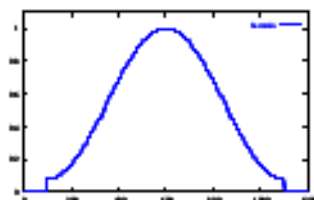
▶ 哈宁窗 (Hanning) :

$$W_{HN} = \begin{cases} 0.5 - 0.5 \cos(2\pi n / (N-1)) & (0 \leq n < N-1) \\ 0 & (\text{Other}) \end{cases}$$

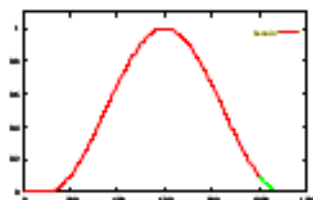
时域上的加窗效果



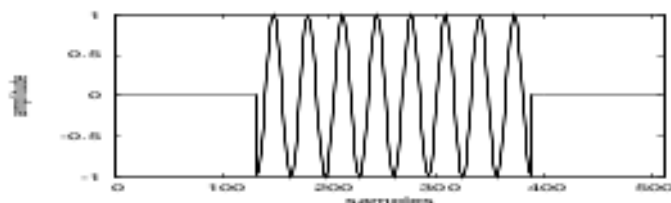
Rectangular



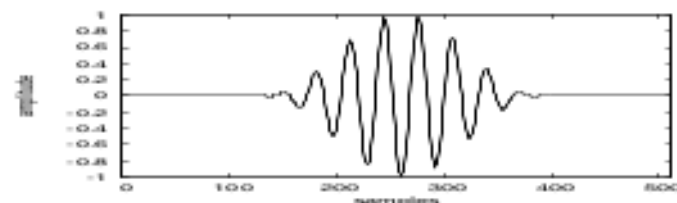
Hamming



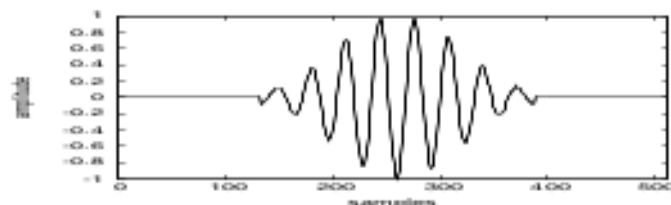
Hanning



(a) Rectangular window



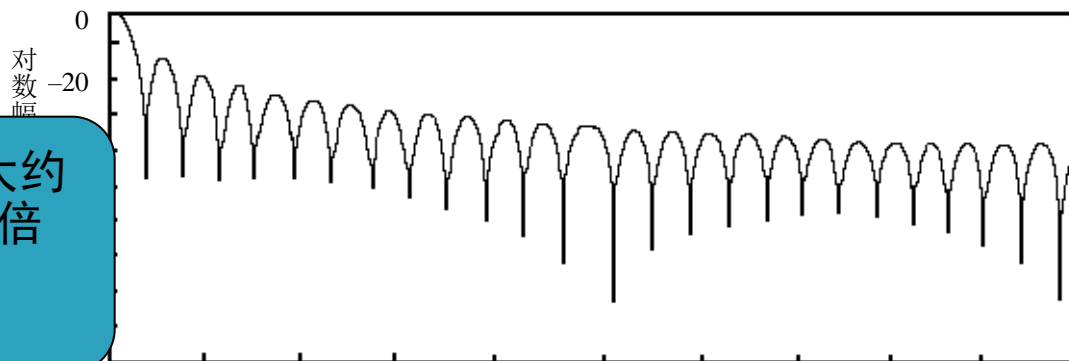
(b) Hanning window



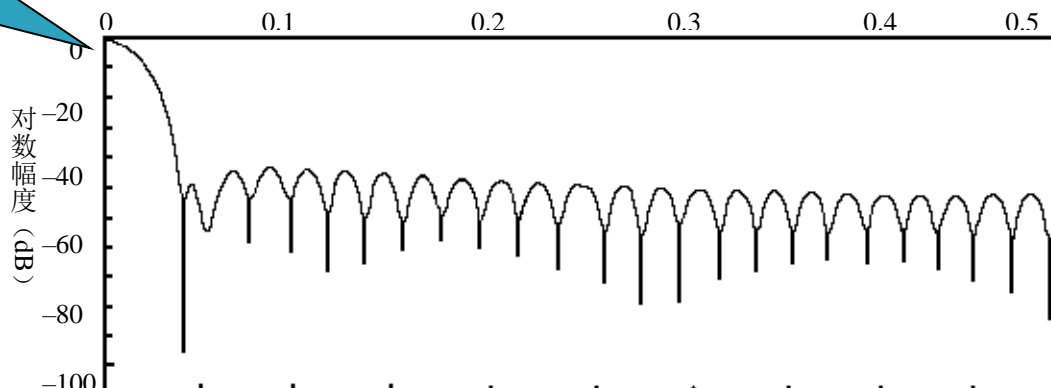
(c) Hamming window

典型的窗函数

- ▶ 若把窗函数理解为某个滤波器的单位冲激响应。则可以比较它们的频率响应特性。



方窗



汉明窗

汉明窗的带宽大约是矩形窗的两倍

汉明窗能更好地保留原语音信号的频率特性，使用最广泛。

短时傅里叶变换(STFT)

短时加窗信号:

$$x_l[n] = w[n]x[n + lL], 0 \leq n \leq N - 1$$

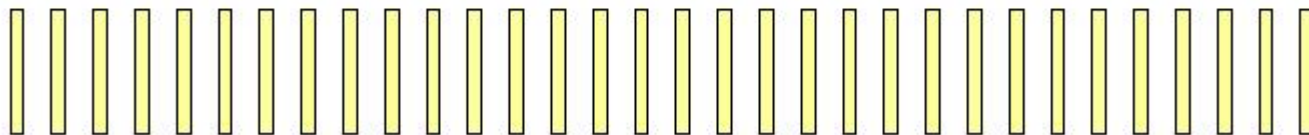
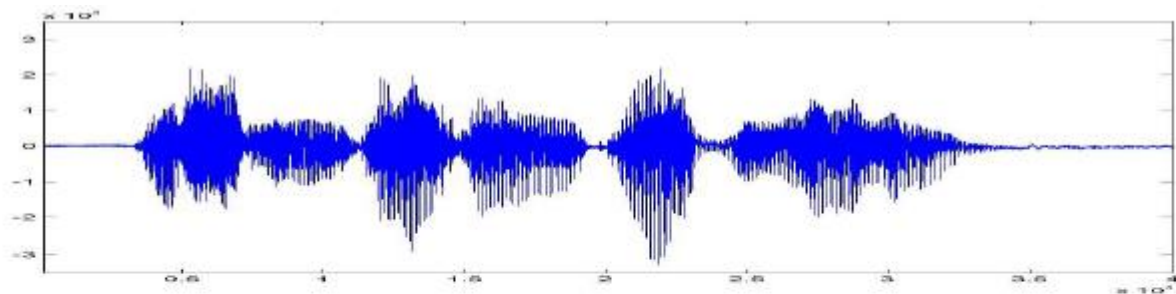
其中 $w[n]$ 是窗函数, N 是窗长, l 是帧索引, L 是帧移。

短时傅立叶变换:

$$X[k, l] = \sum_{n=0}^{N-1} x_l[n] e^{-\frac{j2\pi nk}{K}} = \sum_{n=0}^{N-1} w[n]x[n + lL] e^{-\frac{j2\pi nk}{K}}$$

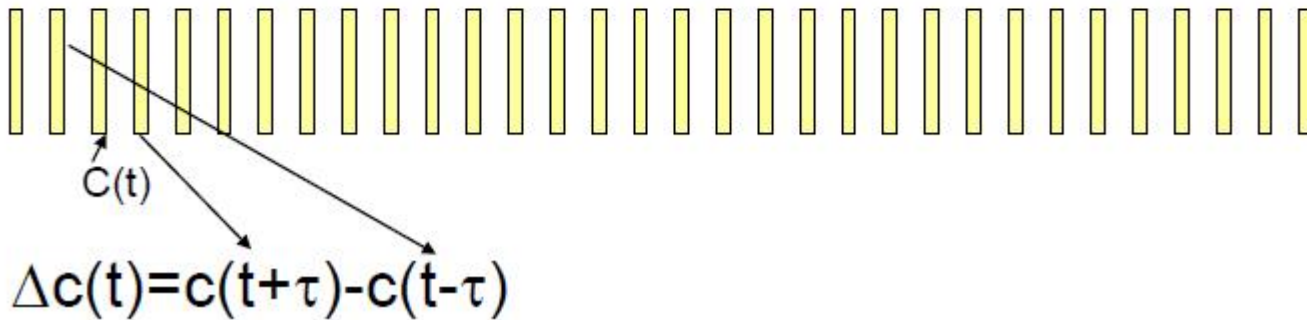
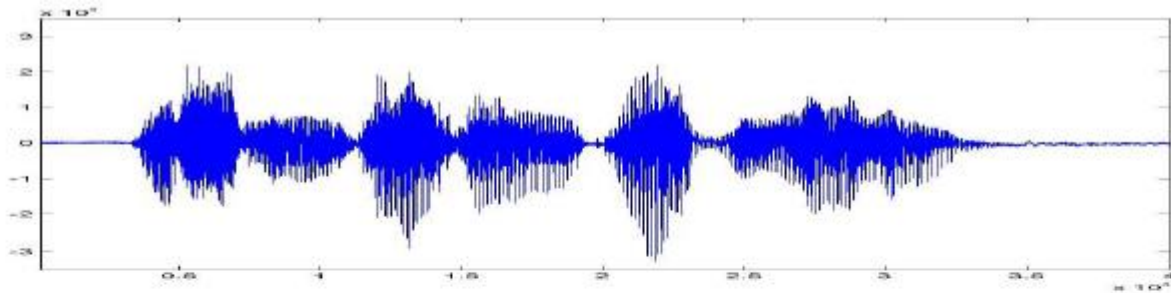
其中 K 是DFT的大小, k 是频率索引。 $X[k, l]$ 建立起时域信号 lL 与频域信号 k 之前的关系, 对于采样率 F_s , 相应的索引对应为时间 lL/F_s 和频率 kF_s/K 。

特征提取过程



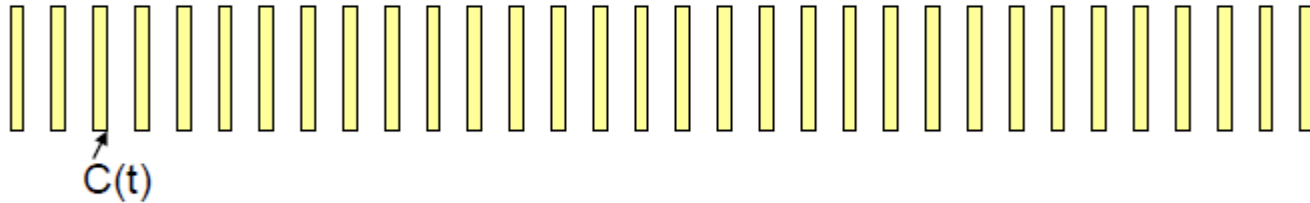
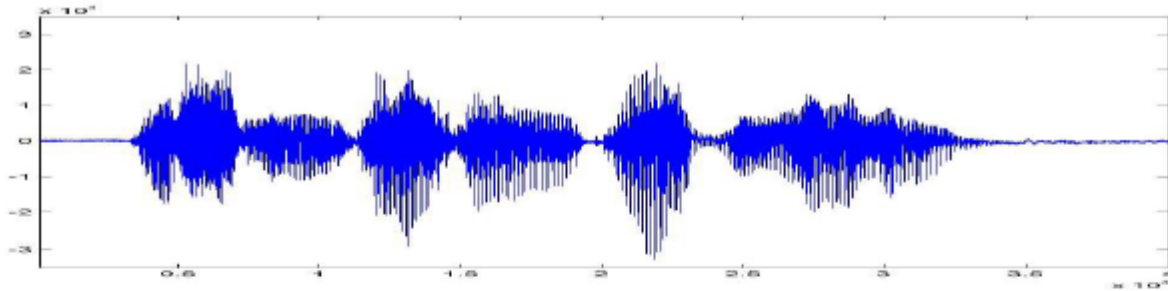
整段语音最后变成一系列特征向量

增加一阶特征(“delta”)



Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides

增加二阶特征(“acceleration”)

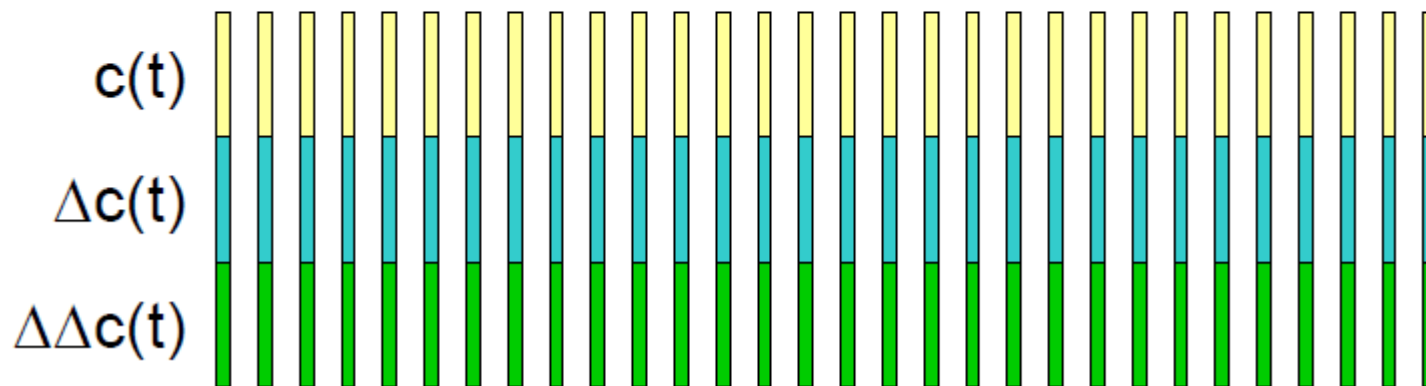
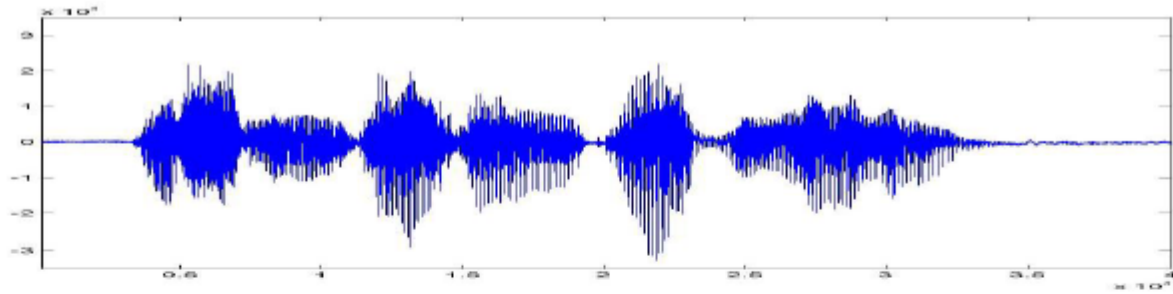


$$\Delta c(t) = c(t+\tau) - c(t-\tau)$$

$$\Delta\Delta c(t) = \Delta c(t+\tau) - \Delta c(t-\tau)$$

Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides

合并特征



Thanks to Prof. Bhiksha Raj and Dr. Ming Li for the contribution of the slides

MFCC特征参数 (39维)

26个滤波器, 12个MFCC系数, 外加短时能量 → 静态特征13维+一阶

特征13维+二阶特征13维

0:	-7.302	-0.601	3.021	-0.667	7.102	-5.344	3.970	-6.495	1.872	-0.641
	4.822	-4.606	50.337	-0.888	0.534	1.002	-0.221	-2.172	0.727	0.558
	0.923	-0.102	-2.401	-0.373	1.710	0.389	0.172	-0.277	-0.143	0.461
	0.208	0.012	-0.346	0.198	0.098	0.421	-0.045	-0.408	-0.131	
1:	-6.488	1.138	1.383	-6.837	2.885	-5.983	4.597	-0.982	3.270	-4.508
	5.388	-3.372	50.404	-0.657	-0.016	0.750	0.548	-0.948	1.106	-0.120
	1.980	0.328	-1.268	0.356	1.714	0.130	0.374	-0.374	-0.335	0.673
	0.299	-0.023	-0.445	-0.287	0.129	0.890	-0.241	-0.799	-0.234	
2:	-12.149	1.201	8.851	1.311	-1.648	-1.391	6.448	-4.635	0.664	-10.713
	2.673	3.328	52.248	-0.145	-0.576	0.411	1.697	-1.745	0.595	-0.835
	1.383	0.171	-0.865	-0.963	-0.331	-0.139	0.514	-0.303	-0.432	0.176
	0.483	-0.323	-0.239	-0.367	0.063	1.193	-0.188	-0.648	-0.279	
3:	-8.162	-1.581	3.855	1.082	6.734	-1.792	2.133	2.476	4.118	-1.944
	7.678	-0.004	50.031	0.611	-0.781	-0.380	2.185	-0.889	0.680	-0.969
	-0.744	0.408	1.281	-1.283	-1.262	-0.519	0.220	0.126	-0.105	-0.420
	0.226	-0.573	0.116	-0.607	-0.109	1.038	-0.192	-0.029	-0.100	

6.常用的语音声学特征

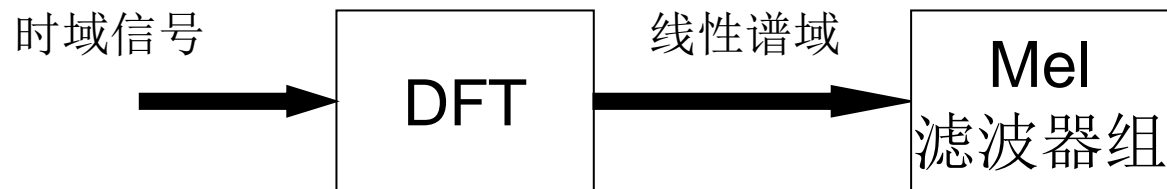
- ▶ Mel频率倒谱系数(MFCC)
- ▶ 滤波器组(FBank)
- ▶ 感知线性预测系数(PLP)

MFCC优势

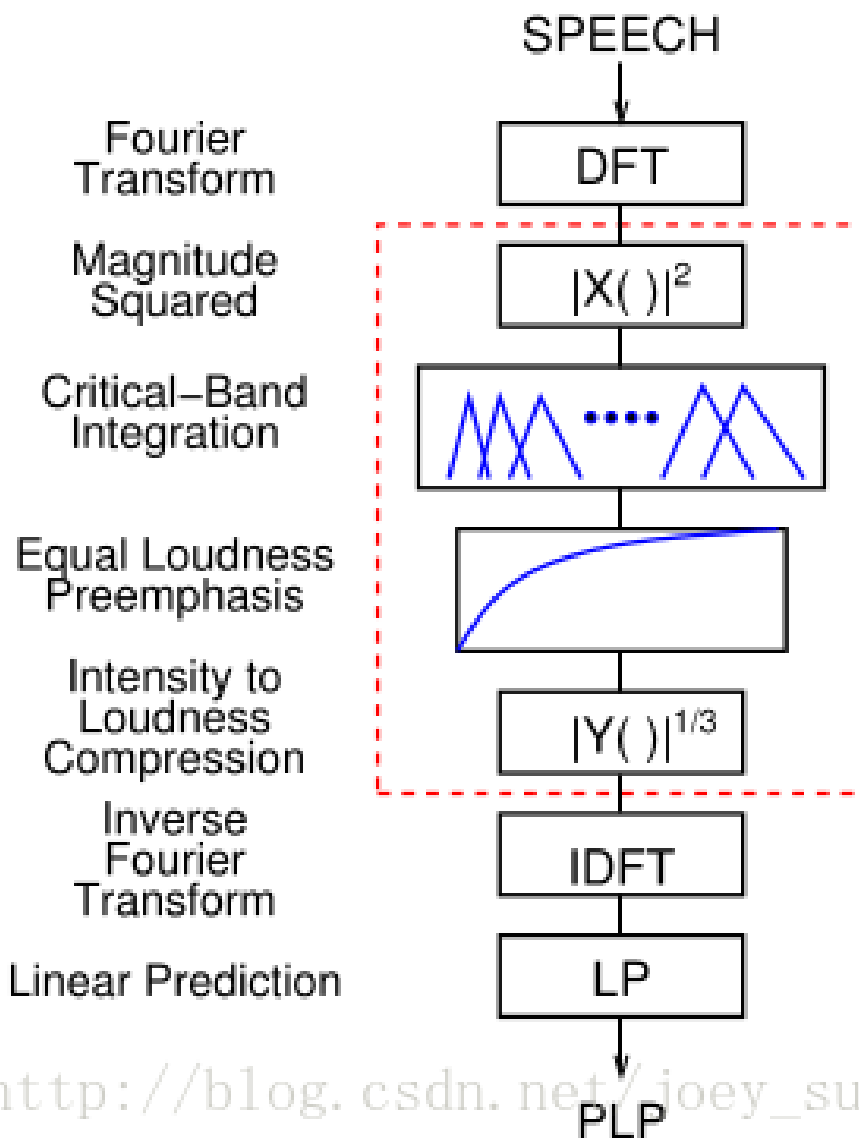
- ▶ 将人耳的听觉感知特性和语音的产生机制相结合。
- ▶ 前12个MFCC通常被用作特征向量(也就是移除F0的信息)。
- ▶ 相对频谱特征有着更小的相关性, 更容易建立模型。
- ▶ 它的表示非常紧凑, 因为这12个特征描述了一帧语音数据中的信息。
- ▶ 对于标准的基于HMM的系统, MFCC在语音识别的性能比滤波器组或者语谱特征更优越。
- ▶ 可惜的是MFCC抵抗噪声的鲁棒性不强。

FBank

- ▶ MFCC的前半部分
- ▶ 特征没有去相关，不满足高斯分布
- ▶ 保留更原始的特征，常用于语音识别的深度学习



感知线性预测系数 (Perceptual Linear Predictive, PLP)



- 利用等响度预加重以及立方根压缩（由感知的结果），而不是MFCC用到的对数压缩；利用线性预测自回归模型获得倒谱系数。
- PLP跟MFCC比较，具有更好的语音识别准确度以及更好的噪声鲁棒性。

语音识别声学特征应包含以下特性

- ▶ 特征应包含区分音素与音素之间的有效信息
 - 良好的时间分辨率（10ms）
 - 良好的频率分辨率（~20 channels）
- ▶ 分离基音频率 F_0 以及它的谐波成分
- ▶ 对不同说话人具有鲁棒性
- ▶ 对噪音或者信道失真具有鲁棒性
- ▶ 有着良好的模式识别特性
 - 低维特征
 - 特征独立

来源：http://blog.csdn.net/joey_su/article/details/36414877

Thank you!

Any questions?