第3章 语音特征提取

洪青阳

纲要

3.1 预处理
3.2 短时傅里叶变换
3.3 听觉特性
3.4 线性预测
3.5 倒谱分析
3.6 常用的声学特征
- 3.6.1 语谱图
- 3.6.2 FBank
- 3.6.3 MFCC
- 3.6.4 PLP

3.7 本章小结

3.1 预处理

（1）预加重

（2）分帧

（3）加窗

3.1 预处理—预加重

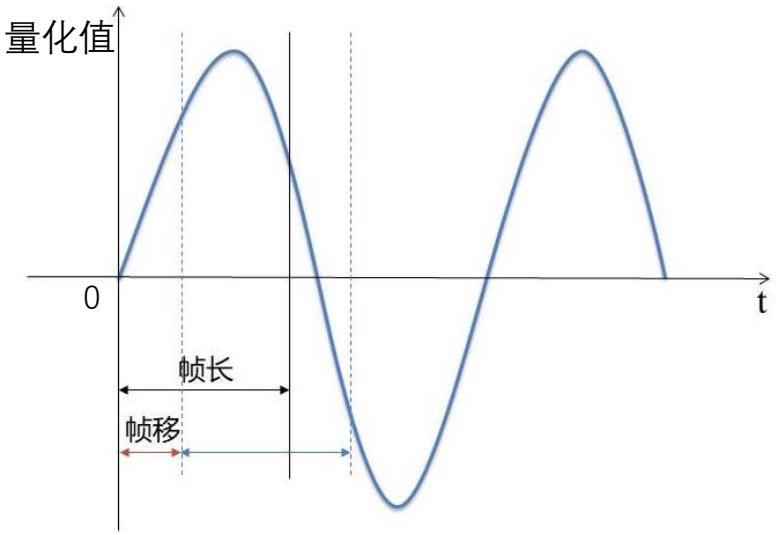
- 语音经发声者的口唇辐射发出，受到唇端辐射抑制，高频能量明显降低，受到的抑制影响越大。

- 可采用**预加重（Pre-emphasis）**补偿语音信号高频部分的振幅。假设输入信号第n个采样点为$x[n]$，预加重公式如下：

$$x'[n] = x[n] - \alpha x[n-1], \alpha = 0.97$$

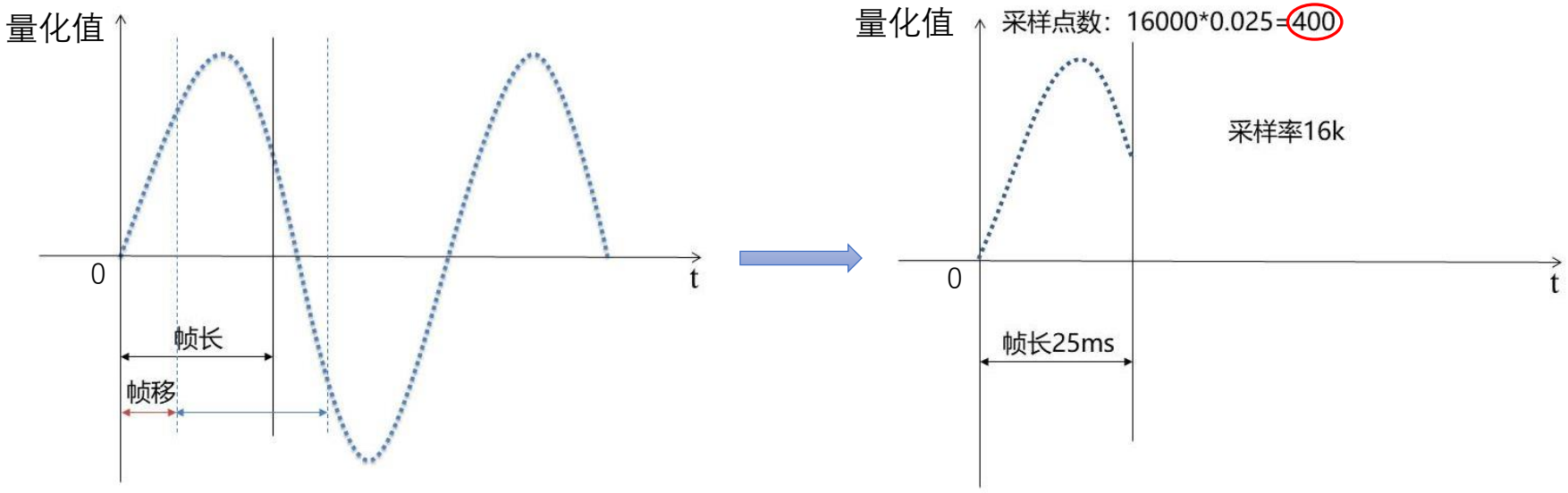
3.1 预处理—分帧

- 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



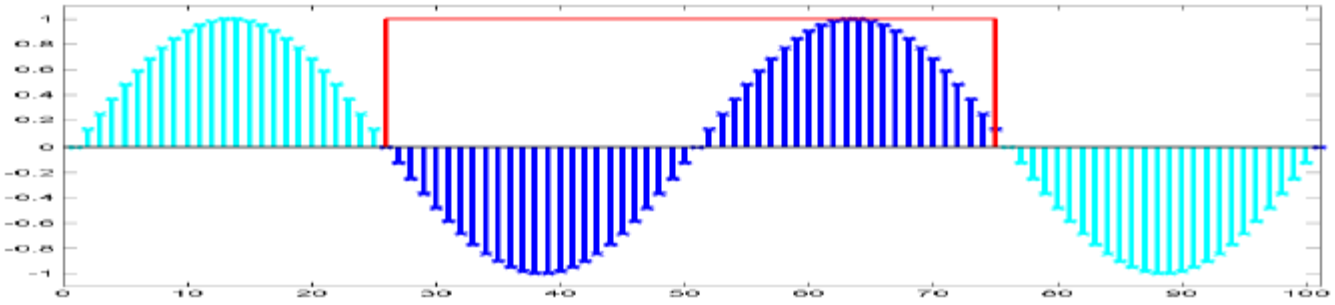
- 语音分帧
  - 每帧10-30ms, 帧间隔10ms

3.1 预处理—分帧

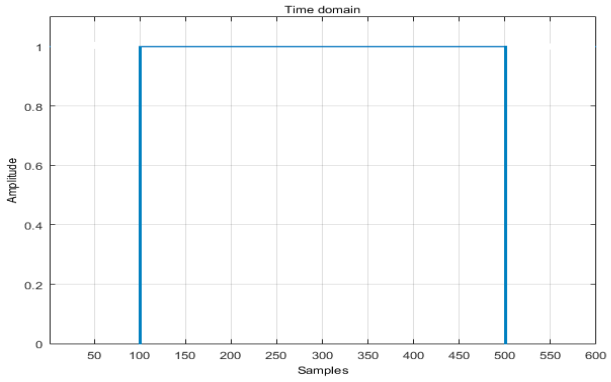


3.1 预处理—加窗

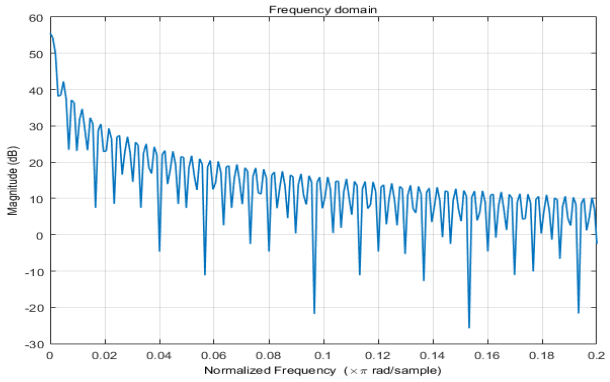
- 分帧方式相当于对语音信号进行了加**矩形窗**的处理。



- 矩形窗在时域上对信号进行有限截断，对应频域的通带较窄，边界处存在多个旁瓣，发生严重的频谱泄露。



（a）矩形窗时域波形



（b）矩形窗频谱函数

3.1 预处理—加窗

- 为减少频谱泄露，通常**对每帧的信号进行其它形式的加窗处理**。常用的窗函数有：汉明（Hamming）窗、汉宁窗（Hanning）、布莱克曼窗（Blackman）等。
- **汉明窗**的窗函数为：

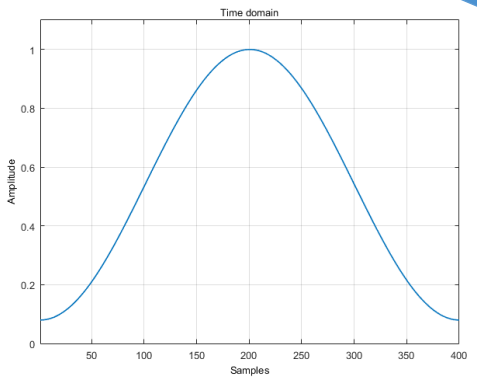
$$w_{ham}[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N\text{-}1}\right)$$

汉明窗能更好地保留原语音信号的频率特性，使用最广泛。

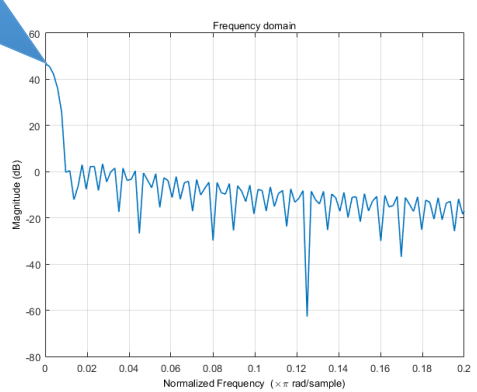
- **汉宁窗**的窗函数为：

$$w_{han}[n] = 0.5[1 - \cos(\frac{2\pi n}{N\text{-}1})]$$

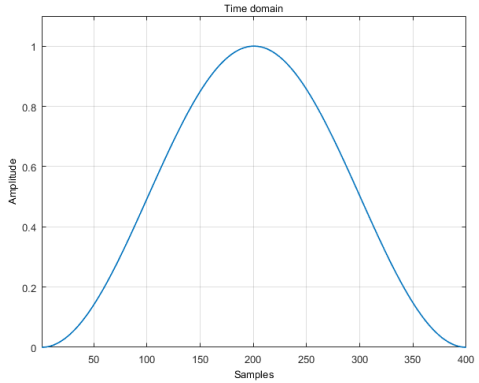
汉明窗的带宽大约是矩形窗的两倍



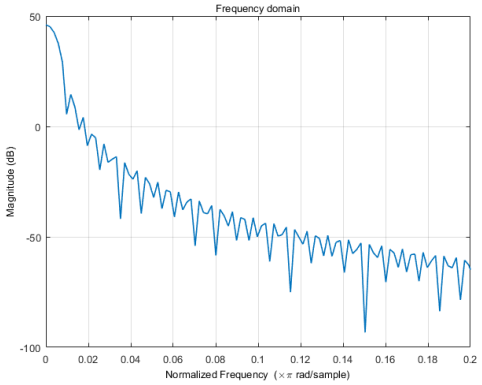
（c）汉明窗时域波形



（d）汉明窗频谱函数



（c）汉宁窗时域波形



（d）汉宁窗频谱函数

3.1 预处理—加窗

- 考虑语音信号的短时平稳性，对每帧语音信号进行加窗处理，得到短时加窗的语音信号$x_l[n]$如下所示：

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

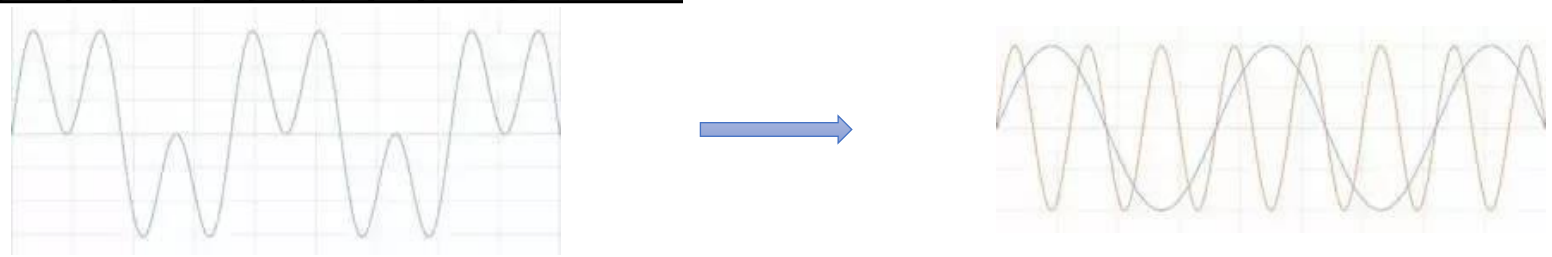
其中，$w[n]$是窗函数，N是窗长，l是帧索引，L是帧移。

3.2 短时傅里叶变换

- 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。

- 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。
- <u>如何分离不同频率信号呢？</u>



3.2 短时傅里叶变换

- 每个频率的信号可以用正弦波表示，采用正弦函数建模。基于欧拉公式，正弦函数对应到统一的指数形式。

$$e^{j\omega n} = \cos(\omega n) + j\sin(\omega n)$$

- 正弦函数具有正交性，即任意两个不同频率的正弦波的乘积，在两者的公共周期内的积分等于零。正交性用复指数运算表示如下：

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

- 基于正弦函数的正交性，语音信号通过相关处理可分离出对应不同频率的正弦信号。

3.2 短时傅里叶变换—DFT

角频率ω = 2πk/K，取值范围是0 ~ 2π；

- 离散傅里叶变换(DFT)的第k个点计算如下：

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

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$K \geq N$。

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

- DFT系数通常是复数的，即

$$e^{-\frac{j2\pi kn}{K}} = \cos\left(\frac{2\pi kn}{K}\right) - j\sin\left(\frac{2\pi kn}{K}\right)$$

则

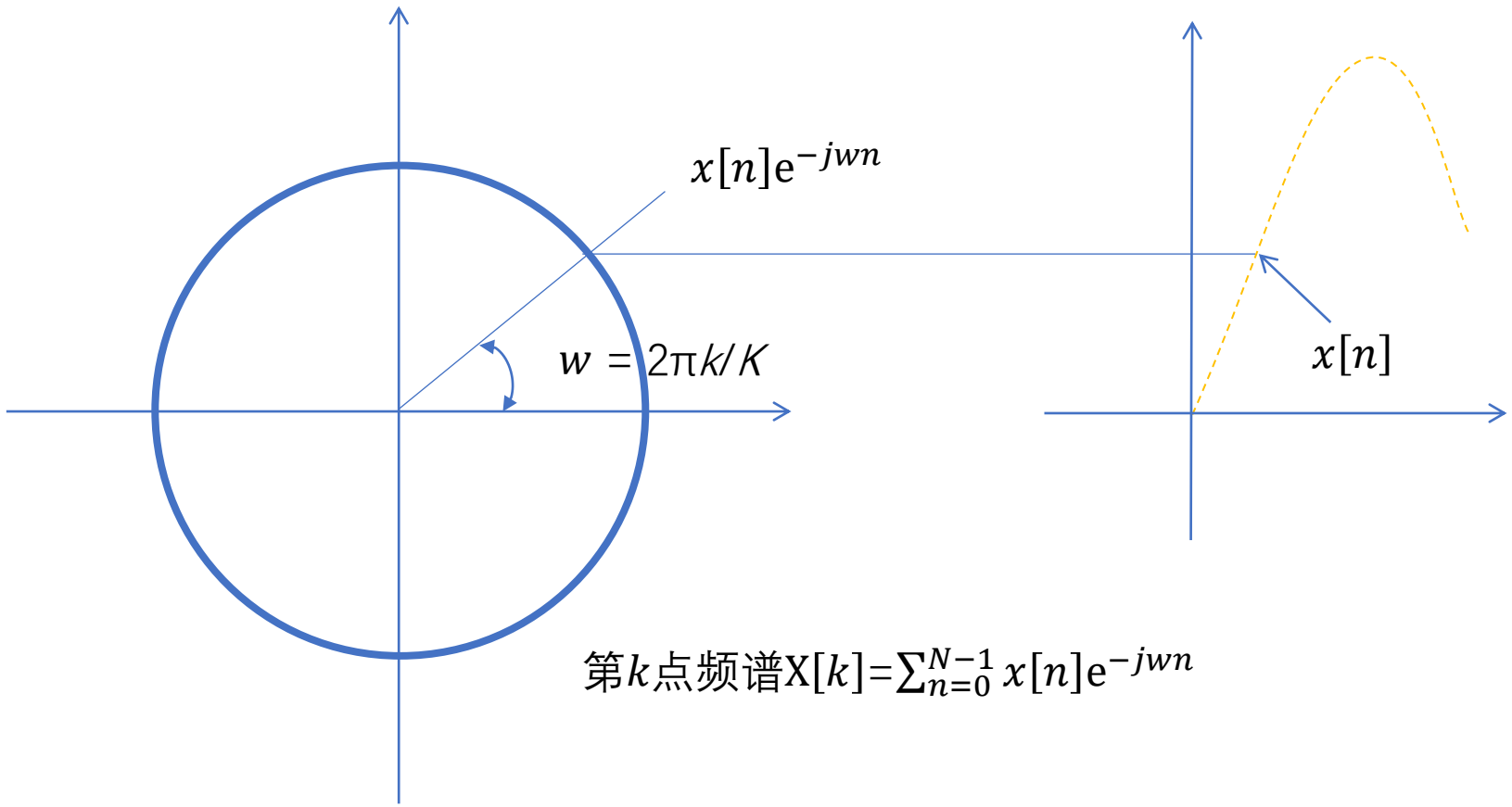
$$X[k] = X_{real}[k] - jX_{imag}[k]$$

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。

其中

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

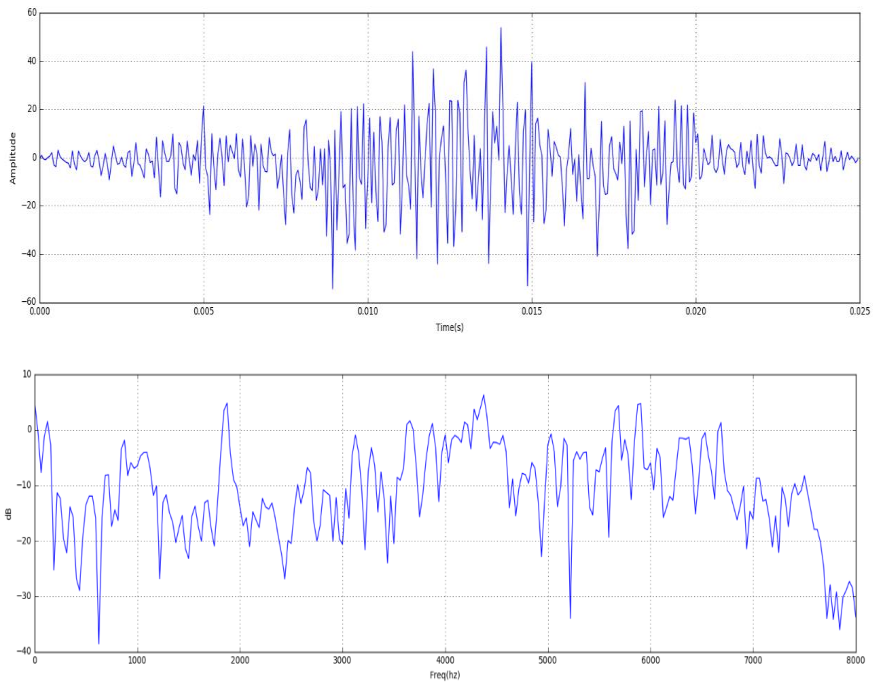
3.2 短时傅里叶变换—DFT



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

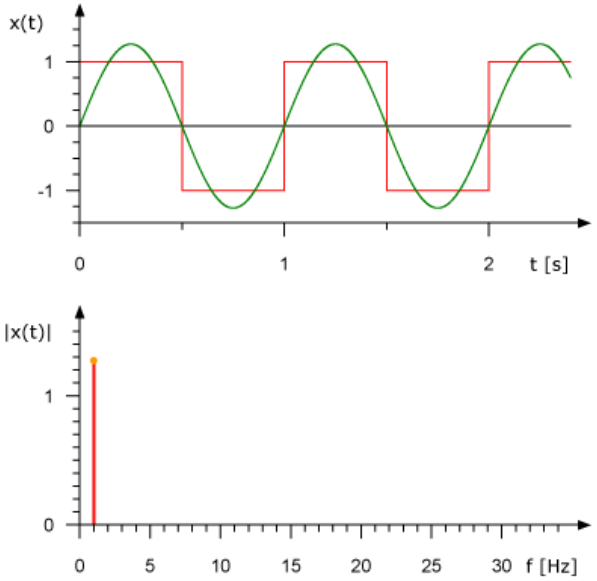
- 时域波形的N个点经DFT后，对应K个点。
- 第0个点代表0Hz，第$K-1$个点代表$(K-1)/K$ *采样率$F_s/2$。K个点在频率轴平均分布。

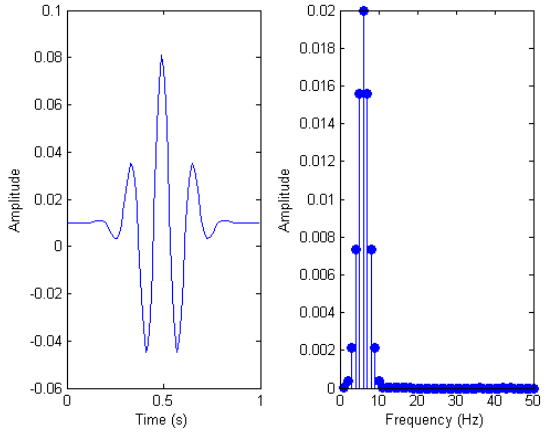
3.2 短时傅里叶变换

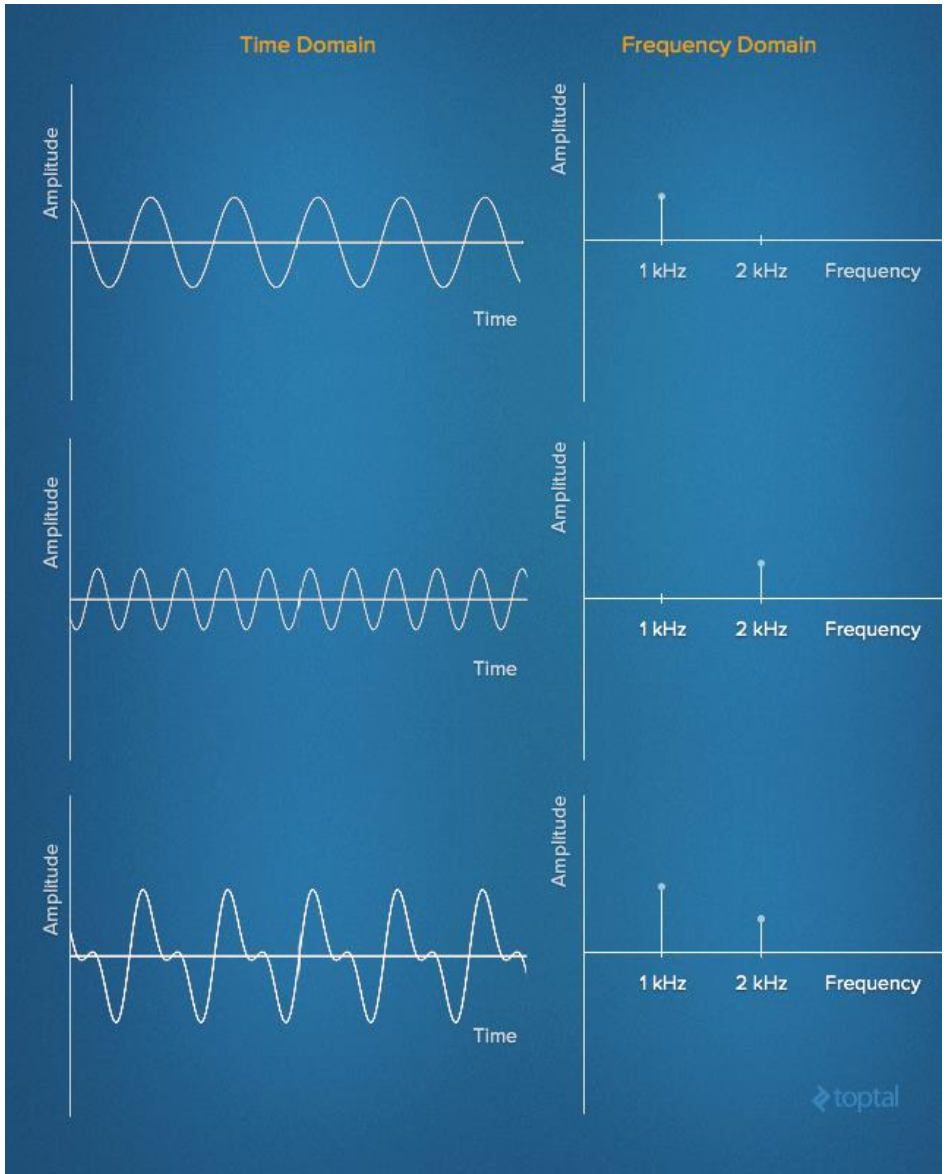


一帧语音的时域波形与频谱图

3.2 短时傅里叶变换


x(t)
1
0
-1
0
1
2
t [s]
|x(t)|
1
0
0
5
10
15
20
25
30
f [Hz]


Amplitude
Time (s)
Amplitude
Frequency (Hz)


Time Domain
Frequency Domain
Amplitude
Time
1 kHz
2 kHz
Frequency
toptal

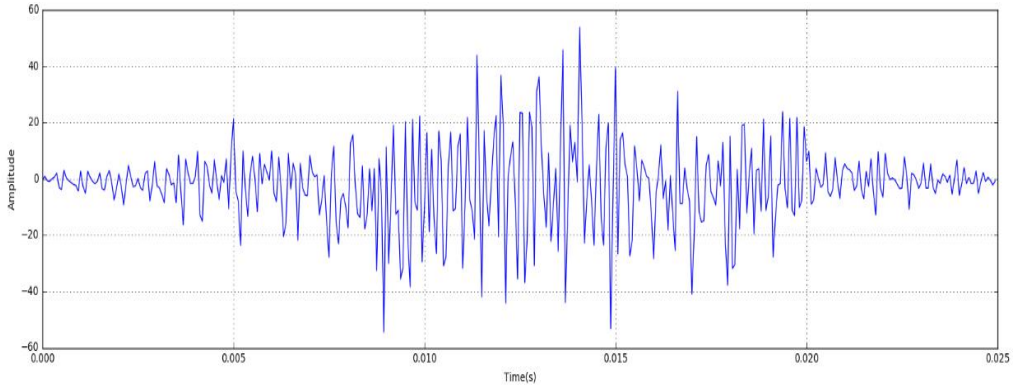
3.2 短时傅里叶变换

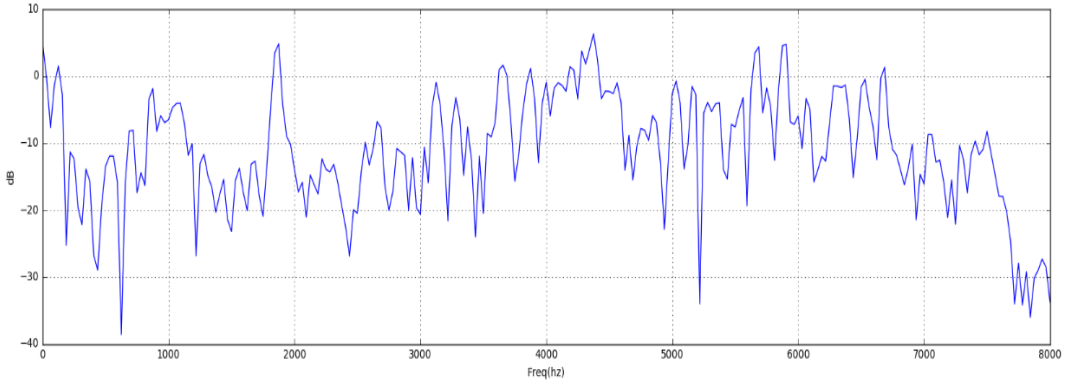
- 振幅频谱用傅里叶系数的幅度表示如下：

$$\mathrm{X}_{\mathrm{magnitude}}[k] = \mathrm{sqrt}(\mathrm{X}_{\mathrm{real}}[k]^2 + \mathrm{X}_{\mathrm{imag}}[k]^2)$$

- 能量频谱用振幅频谱的平方：

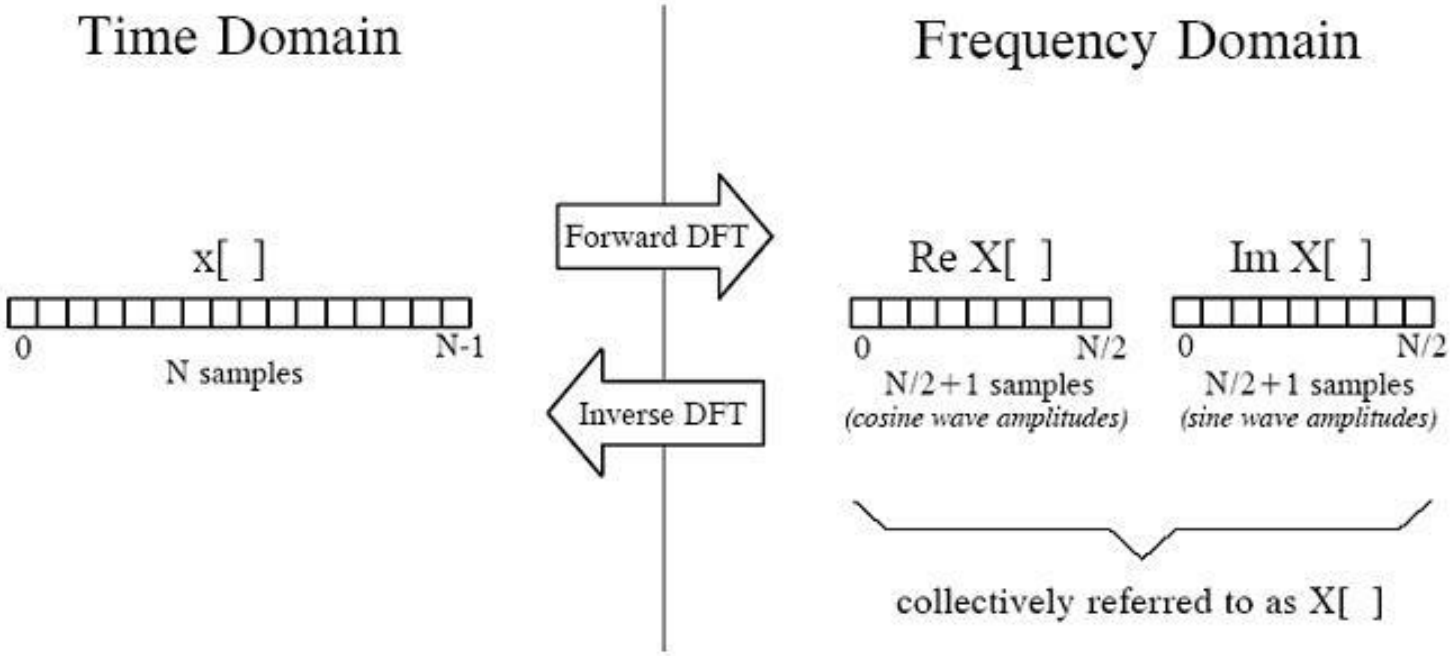
$$\mathrm{X}_{\mathrm{power}}[k] = \mathrm{X}_{\mathrm{real}}[k]^2 + \mathrm{X}_{\mathrm{imag}}[k]^2$$





一帧语音的时域波形与频谱图

3.2 短时傅里叶变换



时域信号可以从DFT恢复得到：

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

3.2 短时傅里叶变换—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，要求窗长N 为2的指数次方。
- 大大简化计算复杂度，在$O(N * log_2 N)$的时间内计算出DFT。
- 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。

3.2 短时傅里叶变换

- 实际应用中，语音信号经过分帧加窗处理，分割成一帧帧的离散序列，可视为采用**短时傅立叶变换（STFT）**：

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

其中K是DFT后的频率点个数，k是频率索引，$0\le k<K$。$\mathrm{X}[k,l]$建立起索引为lL的时域信号与索引为k的频域信号的关系，对于采样率F_s，相应的索引对应为时间lL/F_s和频率$kF_s/(2K)$。

3.3 听觉特性

物理	人类感知	物理解释
声强	响度	人类主观感觉到的声音强弱程度。相同的声强，频率不同时，响度也可能不同。
频率	音调 （音高,pitch）	表示声音的高低，主要由声音的频率决定，同时也与声音强度有关。音高的变化—声调（阴平、阳平、上声、去声）
频谱形状	音色	表示声音的品质，音色与声波的振动波形有关，或者说与声音的频谱结构有关。

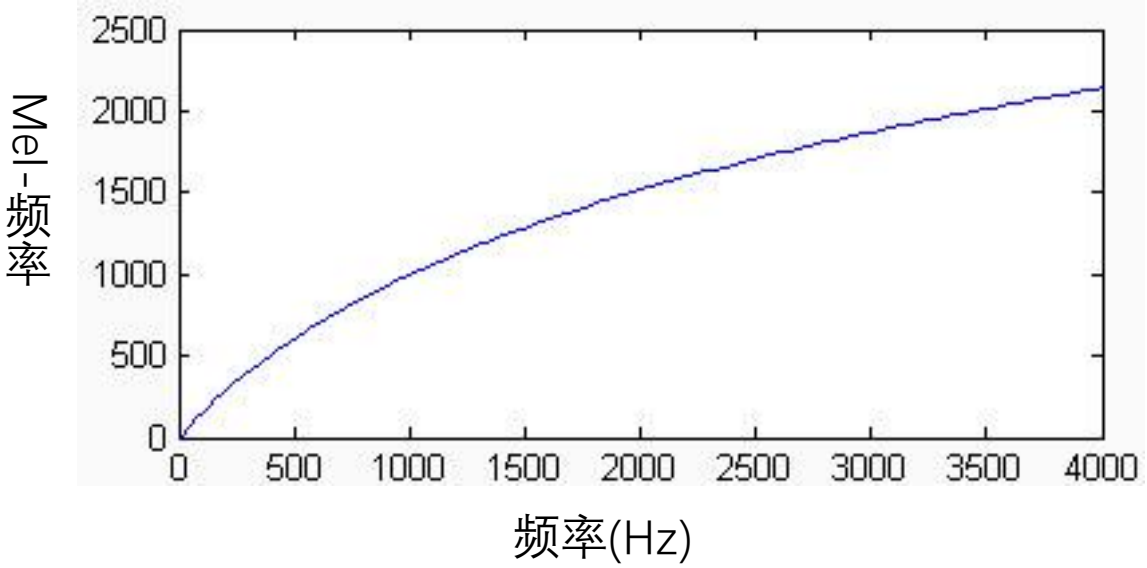
技术术语:

- 音调的单位：Mel频率
- 响度
- Bark频率

3.3 听觉特性—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。



3.3 听觉特性—Mel频率

- Mel频率定义
  - 1Mel—1kHz音调感知程度的1/1000

- Mel频率可以用公式表达如下：

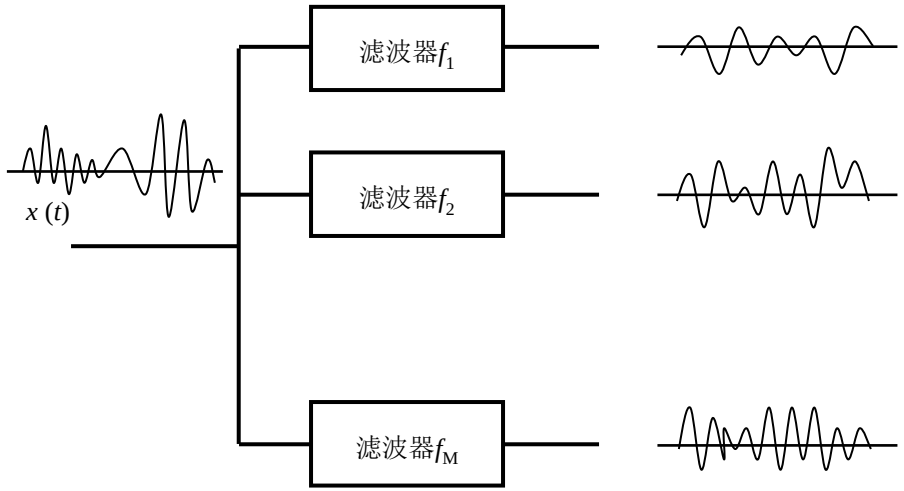
$$\mathrm{Mel}(f) = 2595\lg(1 + f/700)$$

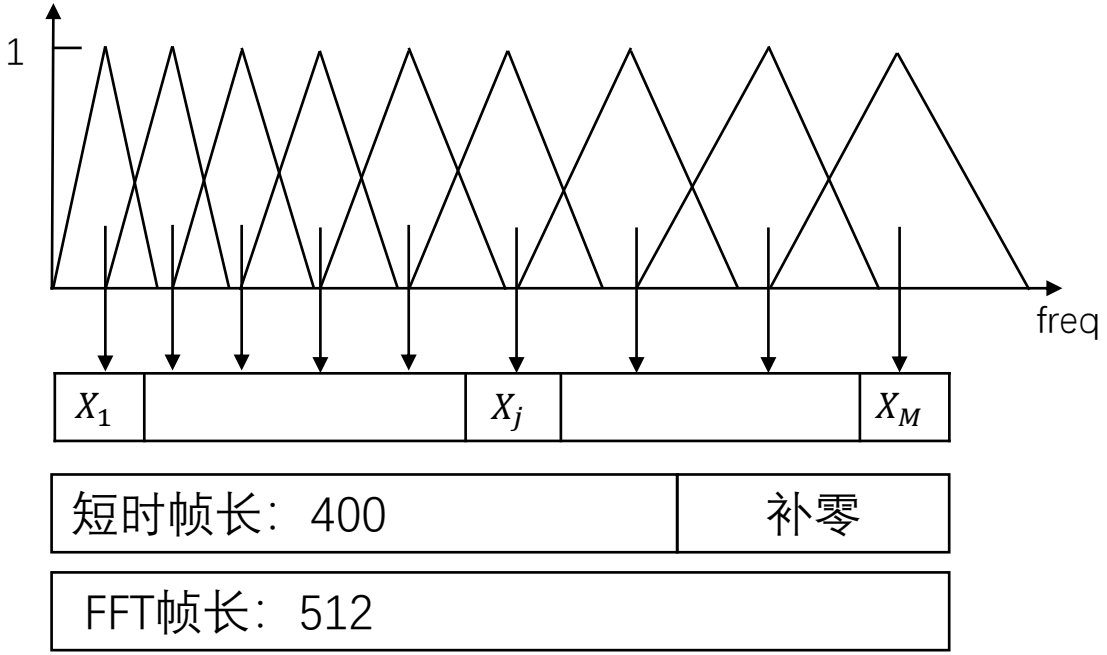
其中lg是以10为底数，即log10。如果以e为底数，则公式为

$$\mathrm{Mel}(f) = 1127\log(1 + f/700)$$

3.3 听觉特性—Mel频率

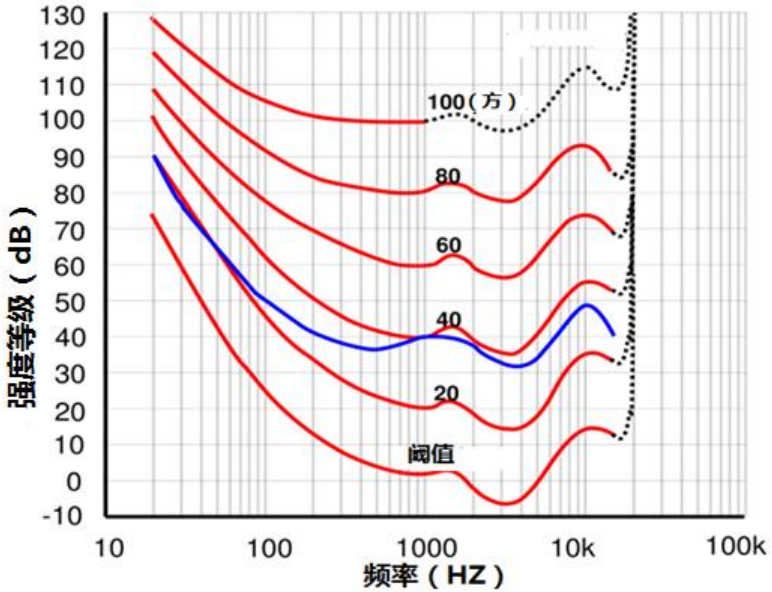
- 研究者根据心理声学实验得到了类似于耳蜗作用的一组滤波器组，模拟人耳对不同频段声音的感知能力。
- 多个带宽不等的三角滤波器组成Mel频率滤波器组，线性频率小于1000Hz的部分为线性间隔，而线性频率大于1000Hz的部分为对数间隔。





3.3 听觉特性—响度

- 声音的响度，是反映人**主观上**对不同频率成分的声音强弱（声强）的物理量，单位为方(phon)，它可以由时变的压力（声压）P来表示，单位为帕斯卡（Pa）。
- 响度与声强、频率的关系可用等响度轮廓曲线表示。



3.3 听觉特性—响度

- 人耳对响度的感知有一个范围，当声音信号低于某个响度时，人耳是无法感知的，这个响度值称为听觉阈值，或称**听阈**；当声音响度强到使人耳感到疼痛时的值，称为痛阈。

- 在实际环境中，当一个较强信号（掩蔽音）存在时，听阈就不等于安静时的阈值，而有所提高。意味着，邻近频率的两个声音信号，弱响度的信号会被强响度的声音信号所掩蔽（Mask），这就是频域掩蔽。

3.3 听觉特性—Bark频率

- 根据听觉频域分辨力和频域掩蔽的特点，Harvey Fletcher提出临界频带的概念，即在某一段频率范围内纯音和噪声功率相等，则该纯音处于刚好能被听到的临界状态。

- 一个临界频带的宽度被称为一个Bark，Bark频率$Z(f)$和线性频率f的对应关系定义如下：

$$Z(f) = 6\ln\left\{\frac{f}{600} + \left[\left(\frac{f}{600}\right)^2 + 1\right]^{\frac{1}{2}}\right\}$$

其中，线性频率f单位为Hz，临界频带$Z(f)$单位为Bark。

3.4 线性预测

- 语音信号$x[n]$可以看成是激励信号$\mathrm{e}[n]$与时变滤波器的单位取样响应$\mathrm{v}[n]$的卷积：

$$x[n] = \mathrm{e}[n] * \mathrm{v}[n]$$

- 如果用一个p阶全极点（AR）系统模拟激励产生语音的过程，设这个AR模型的传递函数为：

$$\mathrm{V}(\mathrm{z}) = \frac{X(z)}{E(z)} = \frac{G}{1 - \sum_{i=1}^{p} a_i z^{-i}}$$

其中，p是阶数，G是增益。

- 通过使线性预测到的采样点值在最小均方误差约束下逼近实际语音采样点值，可以求取一组唯一的预测系数$\{a_i\}$，简称为线性预测编码（Linear Prediction Coding，LPC）系数。

3.5 倒谱分析

- 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)。

- 分离发声源和滤波器
  - 发声源：声源基频F0，动态的声门脉冲
  - 滤波器：声道特征，区分音素(phone)

- 倒谱是通过对对数幅度谱进行逆离散傅立叶变换（IDFT）而得到的。

3.5 倒谱分析

- 倒谱分析，又称为同态滤波，主要采用时频变换，得到对数功率谱，再进行逆变换，分析出倒谱域的倒谱系数。

- 通过傅里叶反变换，得到时域的语音信号倒谱。

$$Z^{-1}\left(\hat{X}[z]\right) = Z^{-1}\left(\hat{E}[z] + \hat{V}[z]\right) = \hat{e}[n] + \hat{v}[z] \approx \hat{x}[n]$$

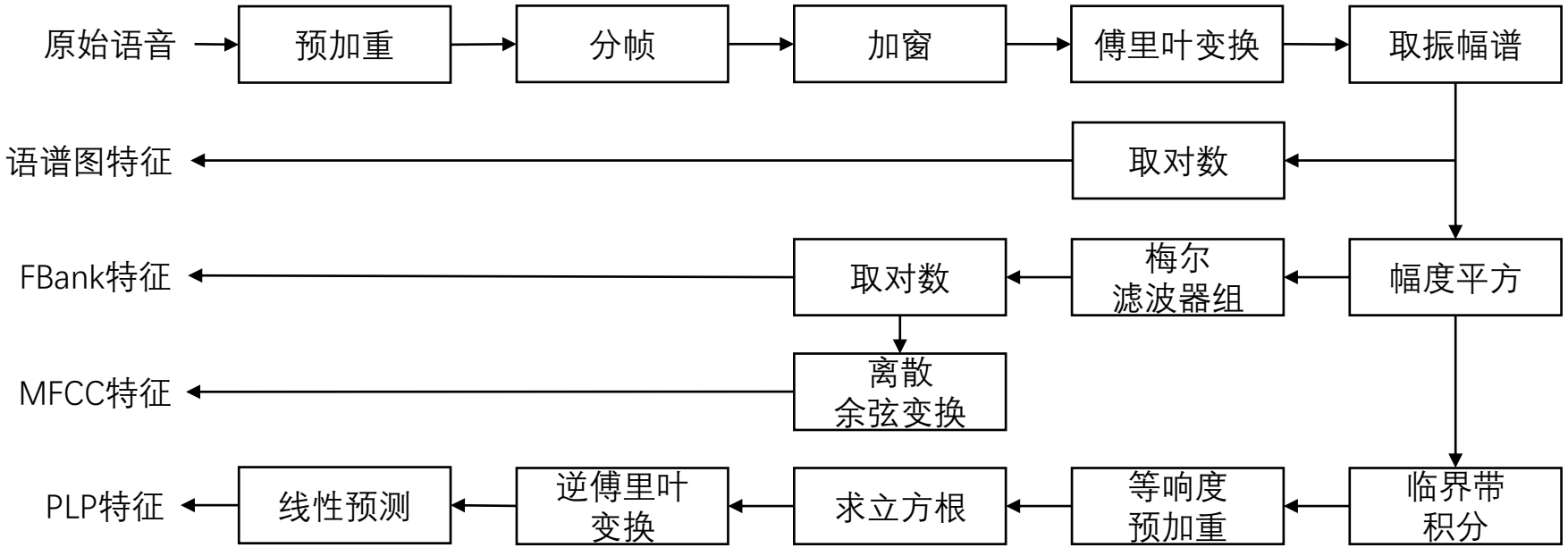
3.5 倒谱分析—DCT

- 在实际应用中，离散余弦变换（DCT）具有最优的去相关性能，能够将信号能量集中到极少数的变换系数上，特别是大多数的自然信号（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。

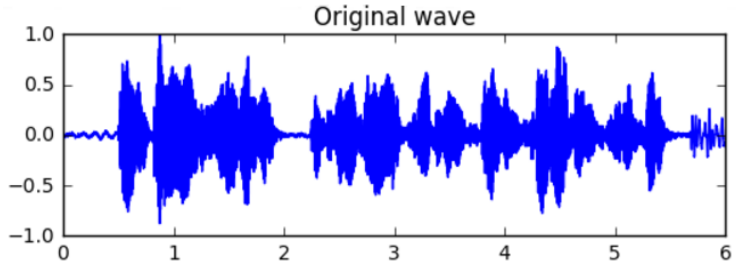
- 一般采用**DCT反变换代替傅里叶反变换**，直接获取低频倒谱系数，对应于包络信息，也就是声道特征。

$$\hat{c}[m] = \sum_{k=1}^{N} \log X[k] \cos\left(\frac{\pi(k-0.5)m}{N}\right), m = 1,2,\cdots,M$$

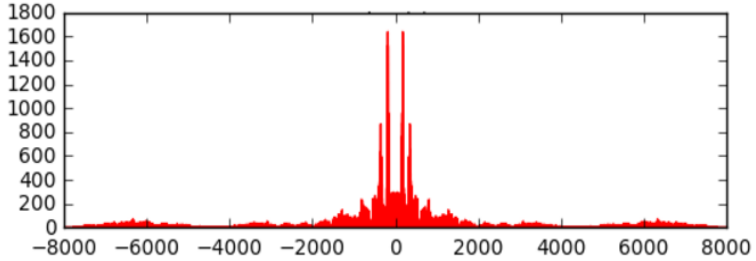
3.6 常用的声学特征


原始语音
预加重
分帧
加窗
傅里叶变换
取振幅谱
语谱图特征
取对数
FBank特征
取对数
梅尔
滤波器组
幅度平方
MFCC特征
离散
余弦变换
PLP特征
线性预测
逆傅里叶
变换
求立方根
等响度
预加重
临界带
积分

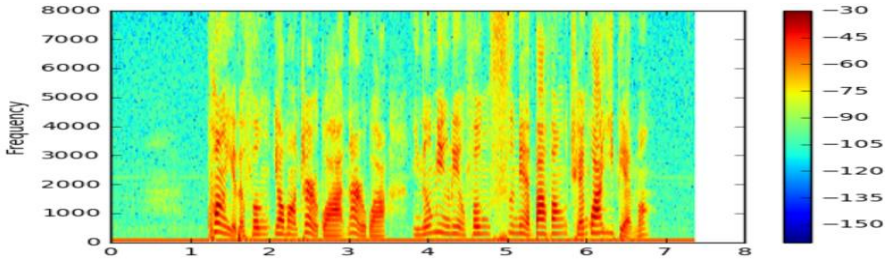
3.6 常用的声学特征—语谱图



(a) 原始语音时域波形（横轴为时间：秒，纵轴为幅值）



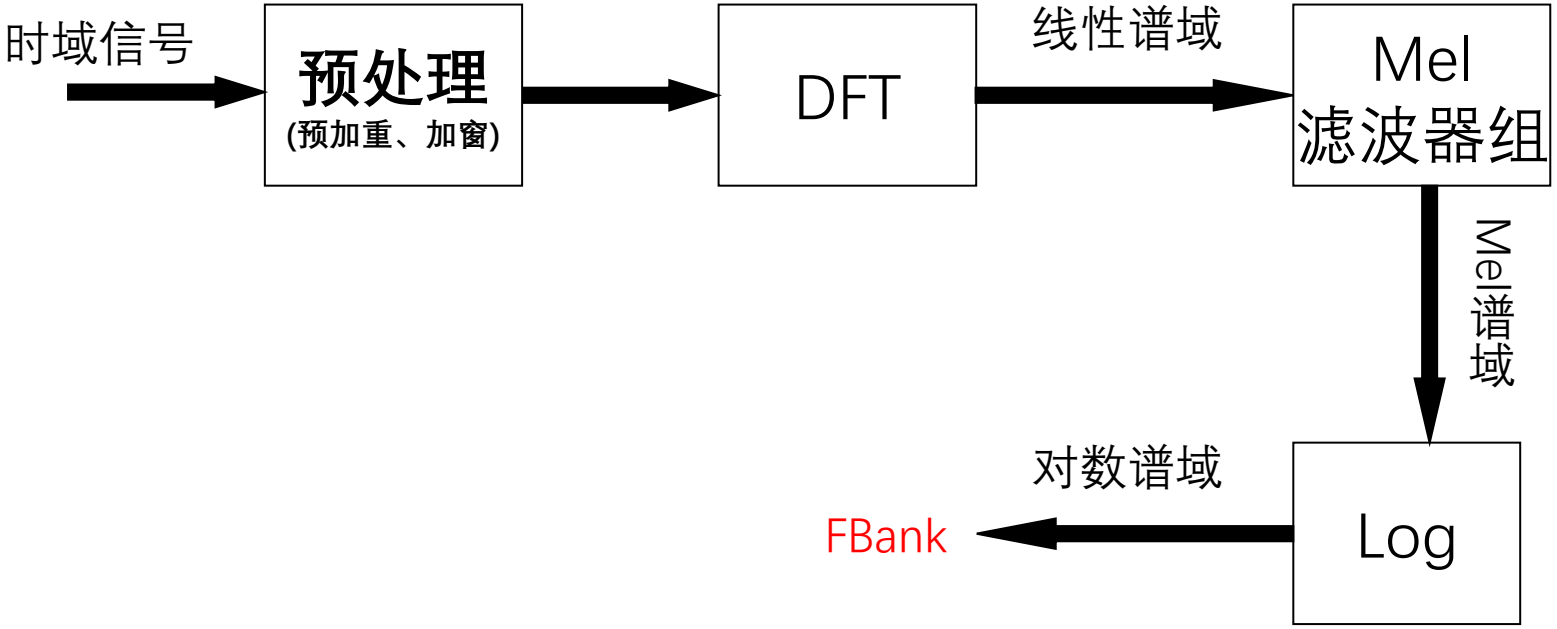
(b) STFT频谱图(横轴为频率：赫兹，纵轴为频谱幅值)



(c) 语谱图(横坐标是时间：秒，纵坐标是频率：赫兹)

（内容为“欢迎访问厦门大学智能语音实验室，我们将带你走进语音的世界。”）

3.6 常用的声学特征—FBank

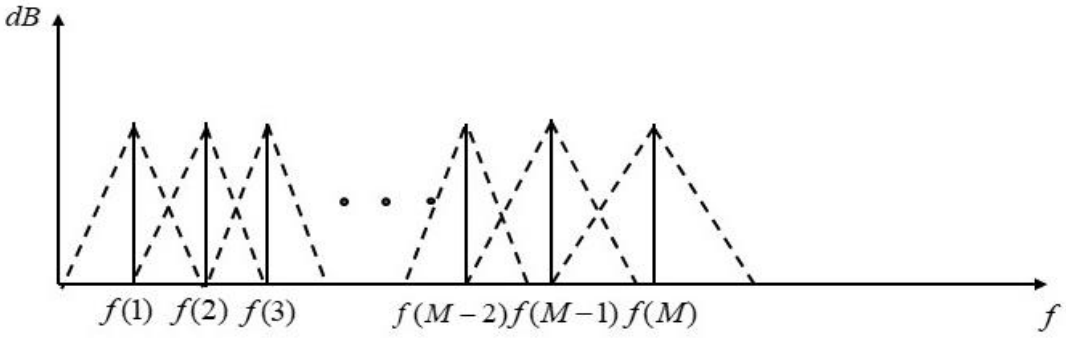


3.6 常用的声学特征—FBank

FBank特征提取流程如下：

（1）将信号进行**预加重**、**分帧**和**加汉明窗**处理，然后进行**短时傅里叶变换(STFT)**得到其频谱；

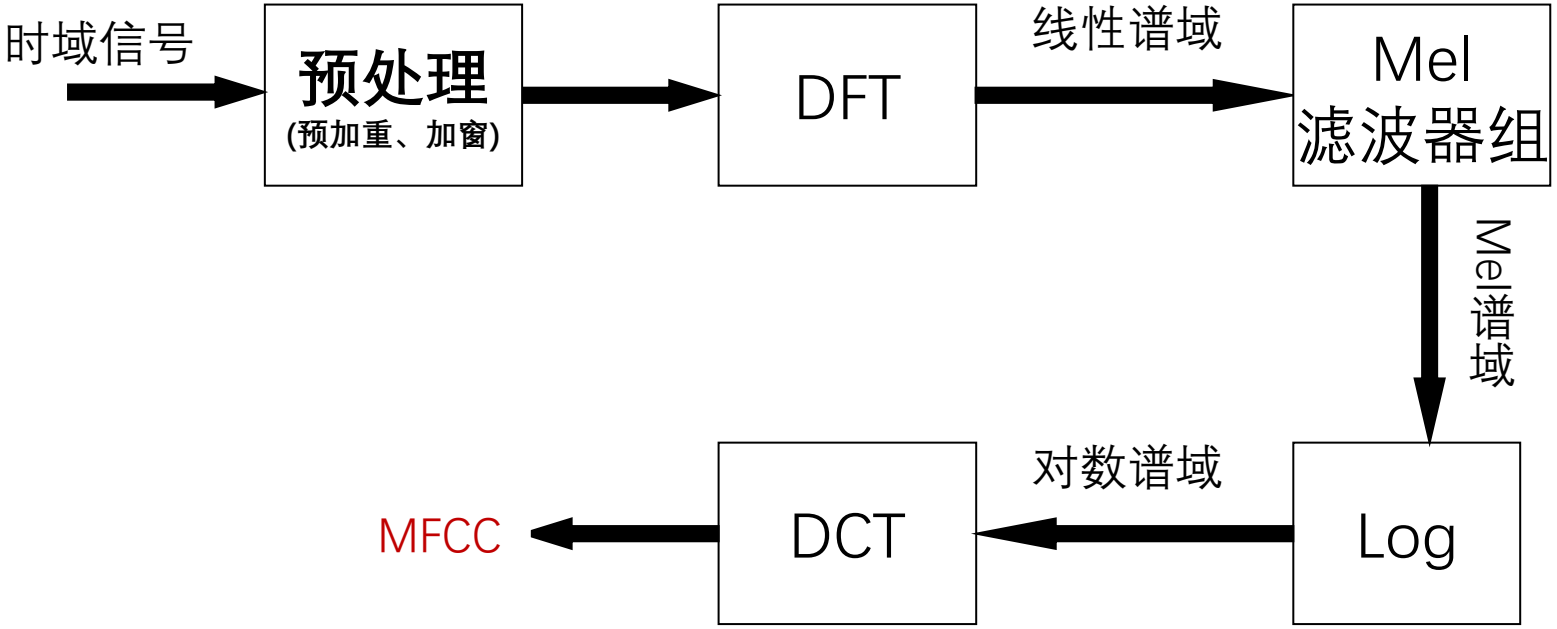
（2）求频谱平方，即能量谱，将每个滤波频带内的能量进行叠加，第k个**滤波器输出功率谱**$X[k]$。



（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。

$$Y_{FBank}[k] = logX[k]$$

3.6 常用的声学特征—MFCC

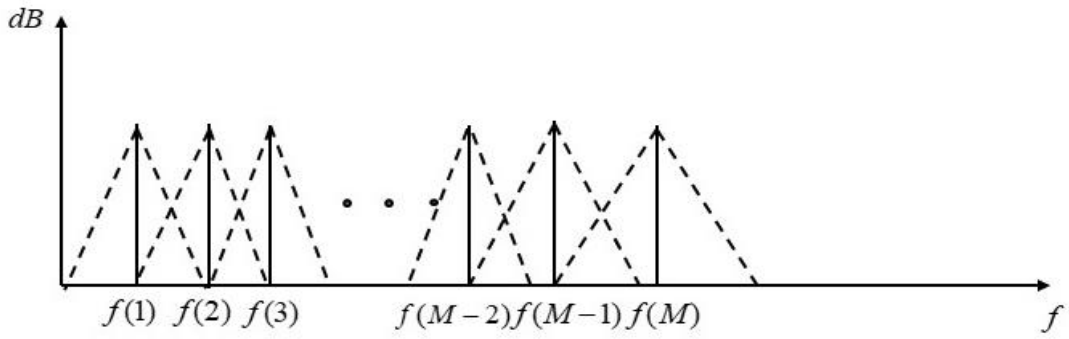


3.6 常用的声学特征—MFCC

MFCC特征提取流程如下：

（1）将信号进行**预加重**、**分帧**和**加汉明窗**处理，然后进行**短时傅里叶变换(STFT)**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(DCT)**，得到L个MFCC系数。

$$C_n = \sum_{k=1}^{M} logX[k]\cos(\frac{\pi(k-0.5)n}{M}), n = 1,2,\cdots,L$$

3.6 常用的声学特征—MFCC

- 语音是时序信号，前后特征有关联；
- 在静态特征基础上，可加上动态特征。

一阶差分：$\Delta c(t) = \frac{c(t+1)-c(t-1)}{2}$

二阶差分：$\Delta\Delta c(t) = \frac{\Delta c(t+1)-\Delta c(t-1)}{2}$

3.6 常用的声学特征—MFCC

例：26个滤波器，12个MFCC系数，外加短时能量（常取对数） ➔
静态特征13维+一阶特征13维+二阶特征13维

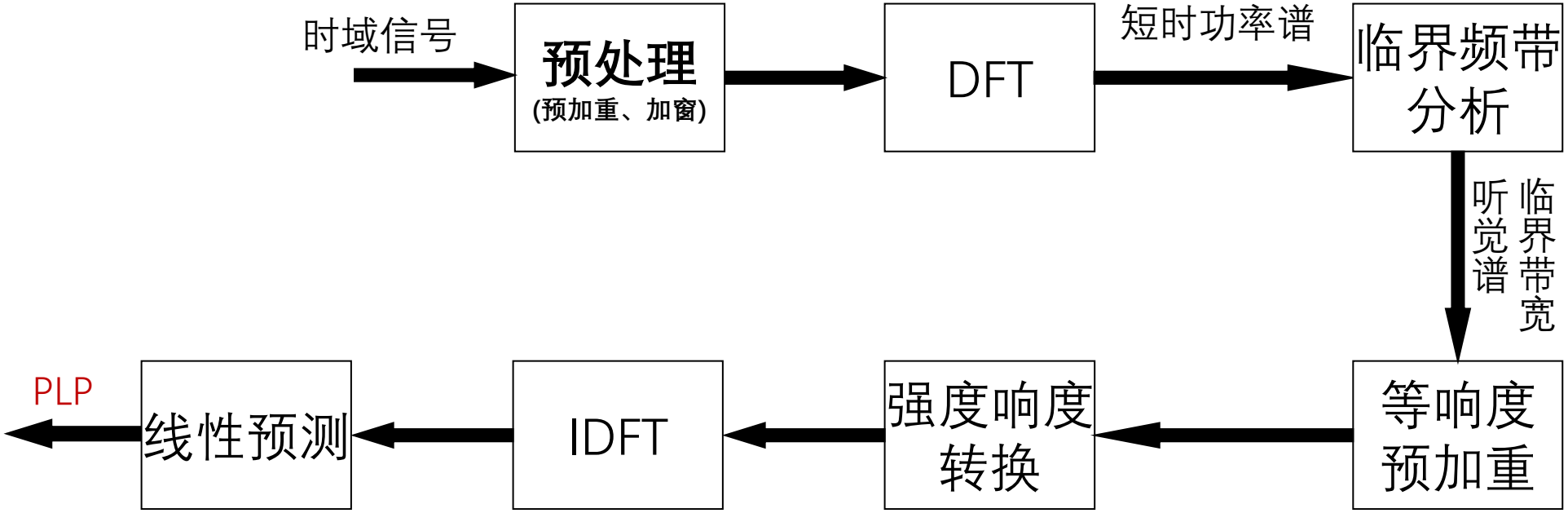
0:	-7.302	-0.601	3.021	-0.667	7.102	-5.344	3.970	-6.495	1.872	-0.641
	4.822	-4.606	50.337	-0.888	0.534	1.002	-0.221	-2.172	0.727	0.558
	0.923	-0.102	-2.401	-0.373	1.710	0.389	0.172	-0.277	-0.143	0.461
	0.208	0.012	-0.346	0.198	0.098	0.421	-0.045	-0.408	-0.131	
1:	-6.488	1.138	1.383	-6.837	2.885	-5.983	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	

3.6 常用的声学特征—MFCC

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)，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。
- 相对FBank特征有着更小的相关性，更容易建立高斯混合模型(GMM)。
- 可惜的是MFCC抵抗噪声的鲁棒性不强。

3.6 常用的声学特征—PLP



- 利用**等响度预加重**以及强度-响度转换（**立方根压缩**），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**。

3.6 常用的声学特征—PLP

PLP特征提取流程如下：

（1）DFT频谱分析

语音信号经过预加重、分帧、加窗、离散傅里叶变换后，取短时语音频谱的实部和虚部的平方和，得到短时功率谱$X_{\text{power}}[k]$。

（2）临界频带分析

将转换得到Bark频率$Z(f)$，一共划分17 个Bark频带。这17个频带中每个频带内的短时功率谱与函数$\psi(Z-Z_0(k))$相乘，求和后得到临界带宽听觉谱$\theta(k)$。

（3）等响度预加重

用模拟人耳大约40dB等响曲线$\mathrm{E}[f_0(\mathrm{k})]$对$\theta(k)$ 进行等响度曲线预加重。

（4）强度-响度转换

为了近似模拟声音的强度与人耳感受的响度间的非线性关系，进行强度-响度转换（立方根压缩）。

（5）离散傅里叶反变换(IDFT)

进行强度-响度转换之后需要进行离散傅里叶反变换。

（6）线性预测(LP)

经过离散傅里叶反变换后，用德宾算法计算12阶全极点模型，并求出16 阶倒谱系数，即为PLP 特征参数。

3.7 本章小结

- 详细介绍了语音特征提取的基本原理，包括预处理、短时傅里叶变换（STFT）。

- 常用的声学特征提取方法：
  - 语谱图、FBank、MFCC和PLP都是采用STFT，有着规律的线性分辨率；
  - FBank和MFCC都采用Mel滤波器组；
  - PLP则利用Bark滤波器组进行模拟人耳听觉特性。

- 优点：
  - FBank保留更多原始特征，适用于深度神经网络(DNN)建模；
  - MFCC去相关性较好，适用于GMM建模；
  - PLP抗噪性更强。