数字媒体技术基础

Meng Yang

www.smartllv.com

SUN YAT-SEN University

机器智能与先进计算教育部重点实验室

智能视觉语言学习研究组

目 录

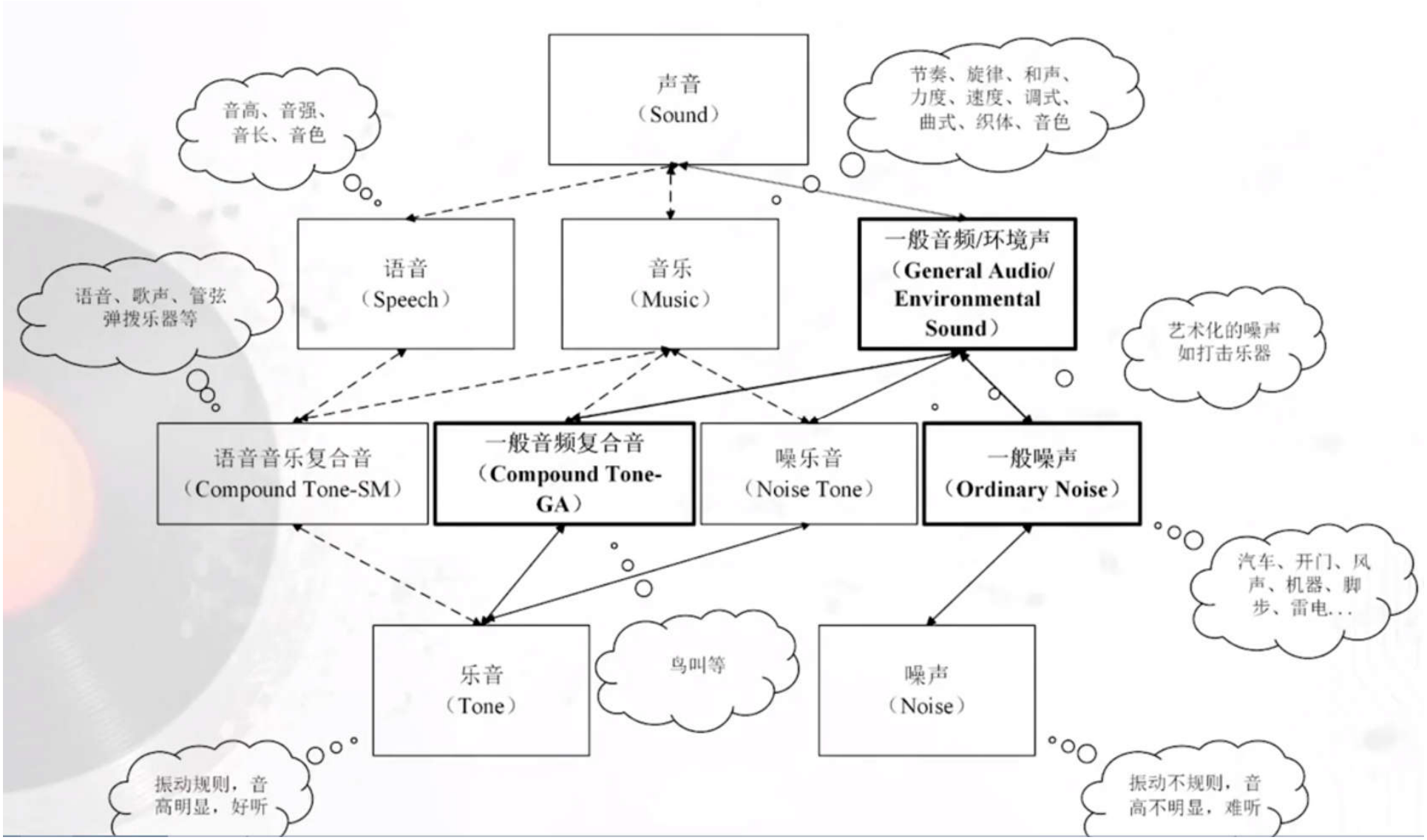
6 音频媒体信息表示

Course Outline

目 录

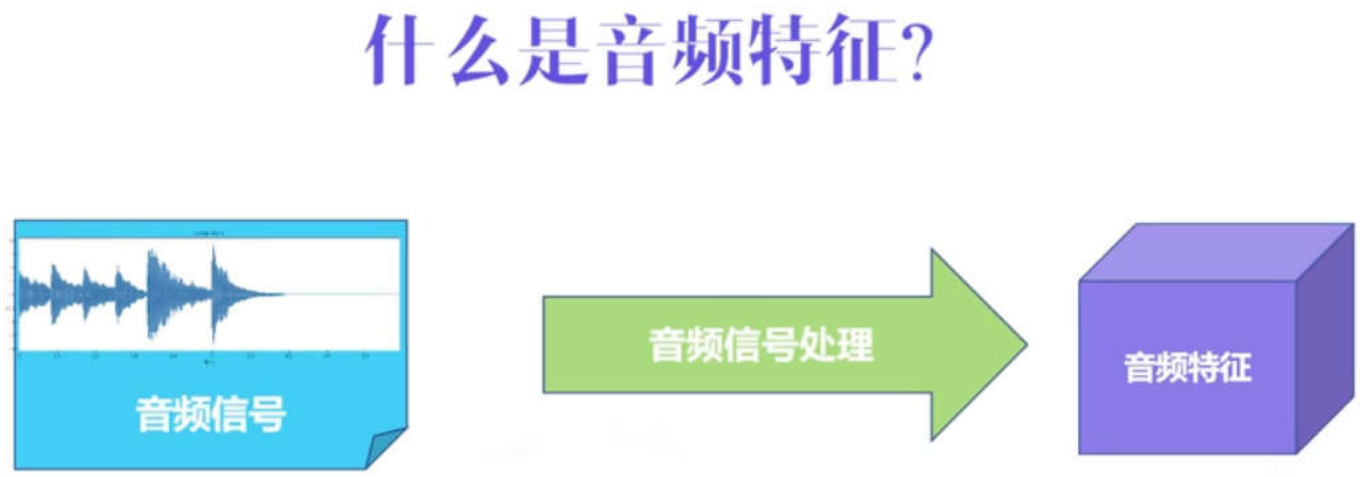
6.3 音频的特征表示

声音的种类


声音
（Sound）
音高、音强、
音长、音色
节奏、旋律、和声、
力度、速度、调式、
曲式、织体、音色
语音
（Speech）
音乐
（Music）
一般音频/环境声
（General Audio/
Environmental
Sound）
语音、歌声、管弦
弹拨乐器等
艺术化的噪声
如打击乐器
语音音乐复合音
（Compound Tone-SM）
一般音频复合音
（Compound Tone-
GA）
噪乐音
（Noise Tone）
一般噪声
（Ordinary Noise）
汽车、开门、风
声、机器、脚
步、雷电...
乐音
（Tone）
鸟叫等
噪声
（Noise）
振动规则，音
高明显，好听
振动不规则，音
高不明显，难听

音频特征

- 绝大部分音频特征最初起源于语音识别中。它们可以精简原始的波形采样信号，从而被利用其它模型中。使算法更容易理解音频中蕴含的语义信息。从20世纪90年代末开始，这些音频特征也被用在音乐信息检索的任务中(比如乐器识别，音符起始点的检测等等)



语音实时识别案例

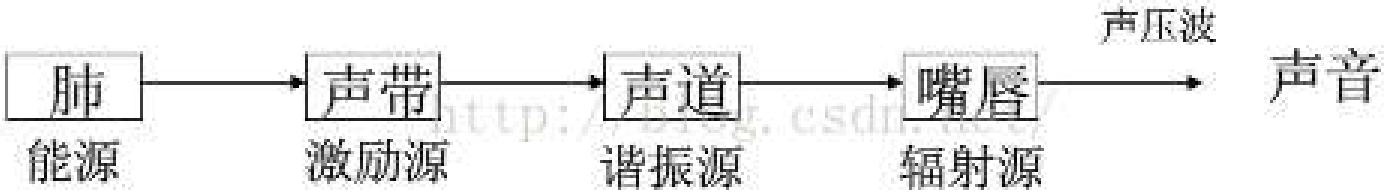
- 会议实时语音识别

- https://haokan.baidu.com/v?vid=4407854321305252835&pd=bjh&fr=bjhauthor&type=video

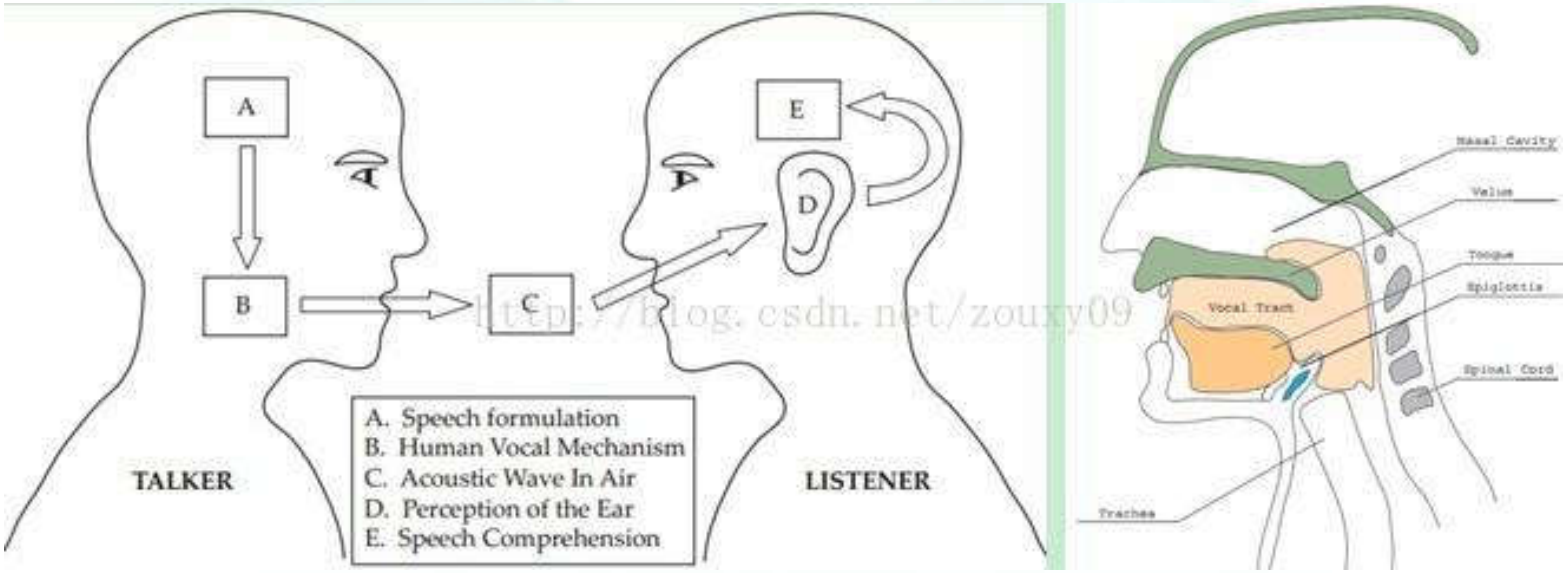
6.3 音频的特征表示

人类声音

- 人通过声道产生声音，声道的shape决定了发出怎样的声音。声道的shape包括舌头，牙齿等。如果我们可以准确的知道这个形状，那么我们就可以对产生的音素phoneme进行准确的描述。声道的形状在语音短时功率谱的包络中显示出来。



人类语音识别


A
B
C
D
E
TALKER
LISTENER
A. Speech formulation
B. Human Vocal Mechanism
C. Acoustic Wave In Air
D. Perception of the Ear
E. Speech Comprehension
Nasal Cavity
Velum
Tongue
Epiglottis
Vocal Tract
Spinal Cord
Trachea

6.3 音频的特征表示

音频的时域和频域

- （1）什么是信号的时域和频域？

  时域和频域是信号的基本性质，用来分析信号的不同角度称为域，一般来说，**时域的表示较为形象与直观，频域分析则更为简练**，剖析问题更为深刻和方便。目前，信号分析的趋势是从时域向频域发展。然而，它们是互相联系，缺一不可，相辅相成的。

问题1

6.3 音频的特征表示

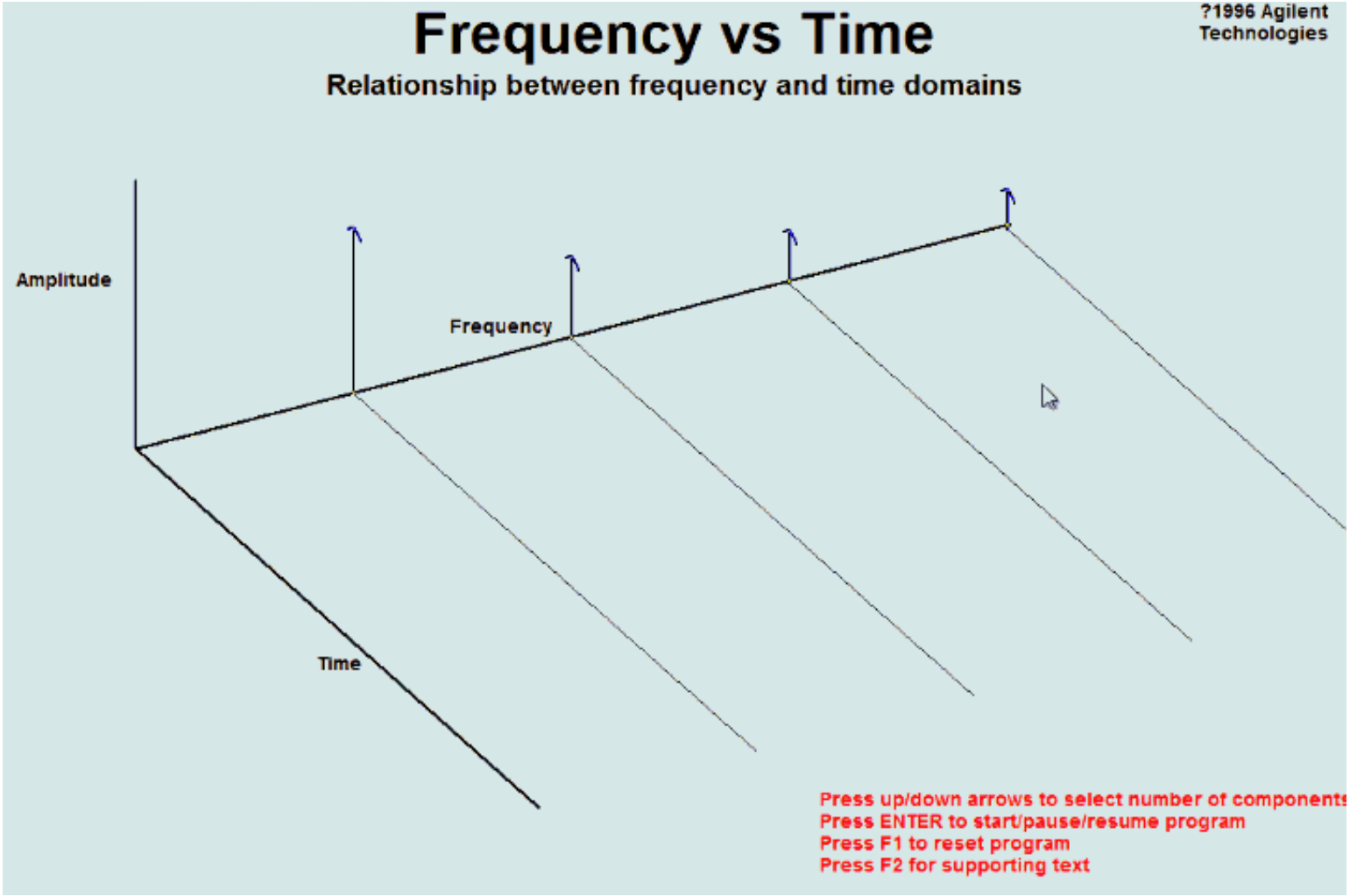
音频的时域和频域

- 时域：自变量是时间，即横轴是时间，纵轴是信号的变化（振幅）。其动态信号x（t）是描述信号在不同时刻取值的函数。

- 频域：自变量是频率，即横轴是频率，纵轴是该频率信号的幅度（振幅），就是指的信号电压大小，也就是通常说的频谱图

6.3 音频的特征表示

时域和频域



6.3 音频的特征表示

时域波形、频域谱线

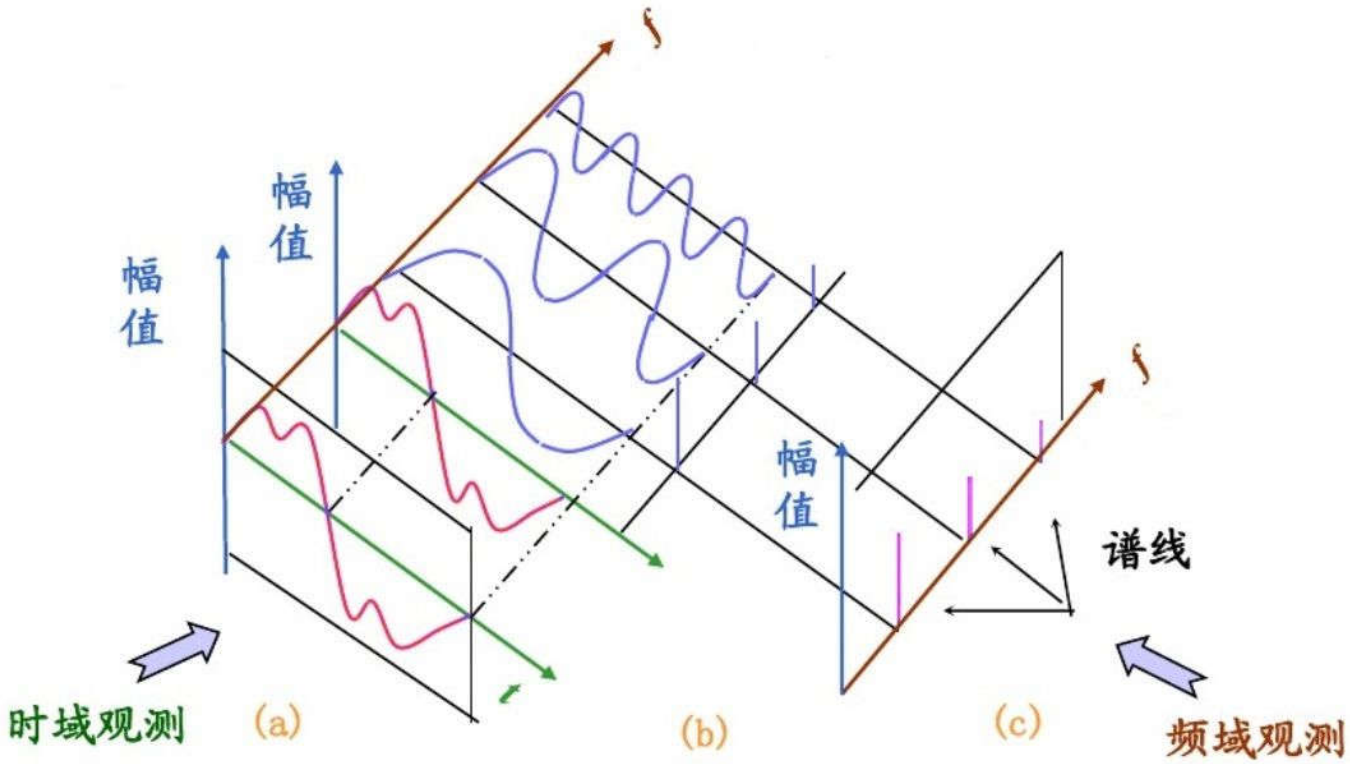


图1-1 波形与频谱

（a）时域波形；（b）时频关系；（c）频域谱线

6.3.1 音频的时域特征表示

- 在时域内提取特征时，通常将研究每个样本的幅度。 我们如何操纵幅度为我们提供了有关信号的某些细节。

$s(k)$ refers to the Amplitdue of the k^{th} sample
K is the frame size, or the number of samples within each frame.
t represents the frame number.

问题2：你觉得有哪些音频特征可以提取？

6.3.1 音频的时域特征表示

振幅包络线

- 振幅包络(Amplitude Envelope)的目的是提取每一帧的最大振幅并将它们串在一起。重要的是要记住振幅代表信号的音量(或响度)。

6.3.1 音频的时域特征表示

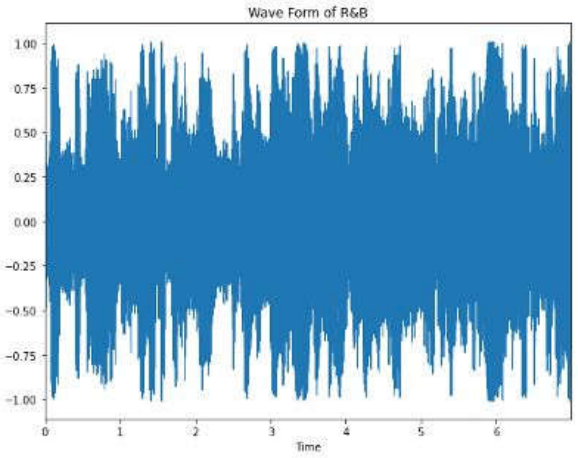
振幅包络线

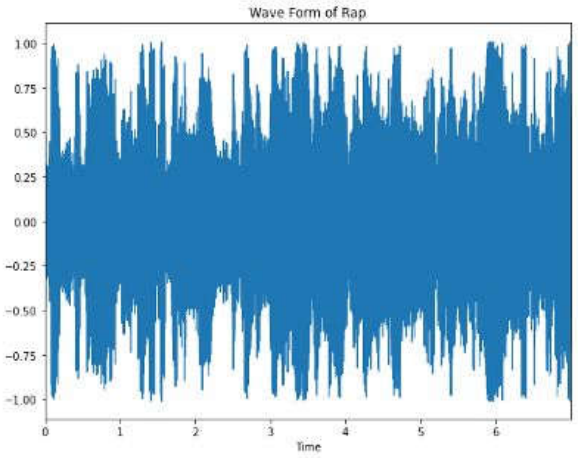
- 首先，我们把信号分解成它的组成窗口，并找出每个窗口内的最大振幅。然后，我们画出每个窗口沿时间的最大振幅。

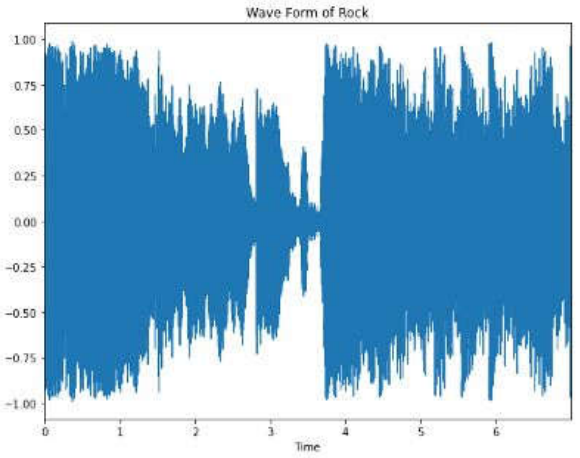
$$\mathrm{AE_t} = \max_{k=tK}^{(t+1)k-1} s(k)$$

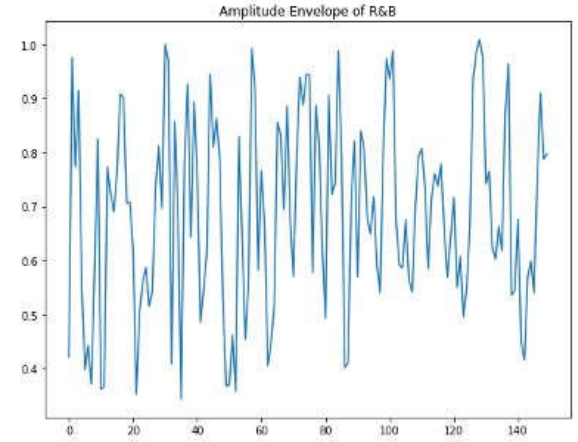
6.3.1 音频的时域特征表示

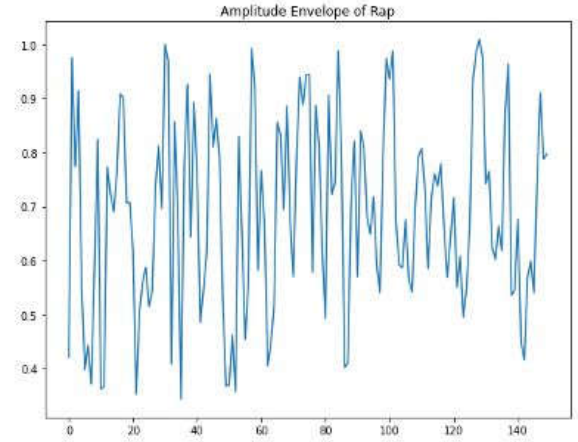
振幅包络线

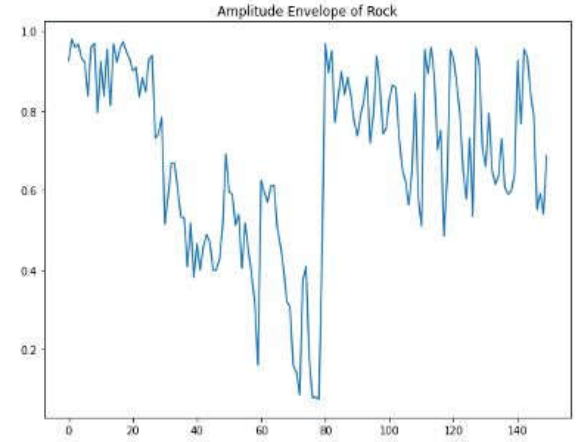












6.3.1 音频的时域特征表示

振幅包络线

- 我们可以将AE用于检测声音是否开始。 在各种语音处理应用程序中，这可能是某人讲话或外部噪音，而在音乐信息检索（MIR）中，这可能是音符或乐器的开始。

6.3.1 音频的时域特征表示

能量的均方根 Root-Mean-Square Energy

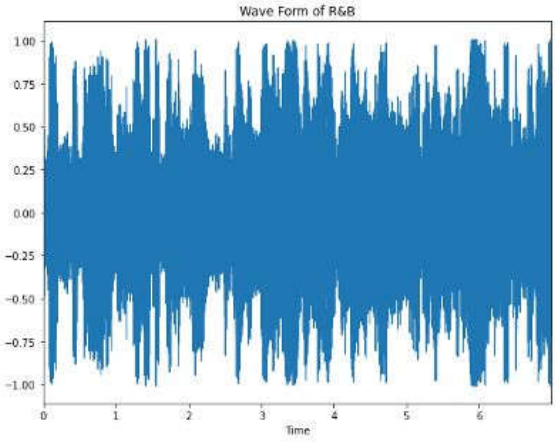
- 均方根（RMS）能量与AE非常相似。 但是，与开始检测相反，它尝试感知响度，该响度可用于事件检测。 此外，它对于异常值的抵抗力要强得多，这意味着如果我们对音频进行分段，就可以更加可靠地检测到新事件（例如新乐器，某人讲话等）。

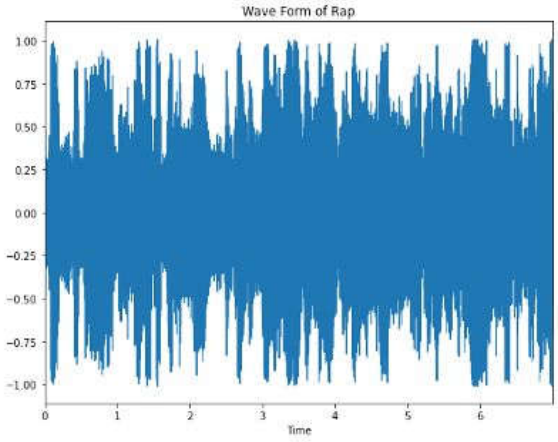
$$RMS_t = \sqrt{\frac{1}{K}\sum_{k=tk}^{(t+1)k-1} s(k)^2}$$

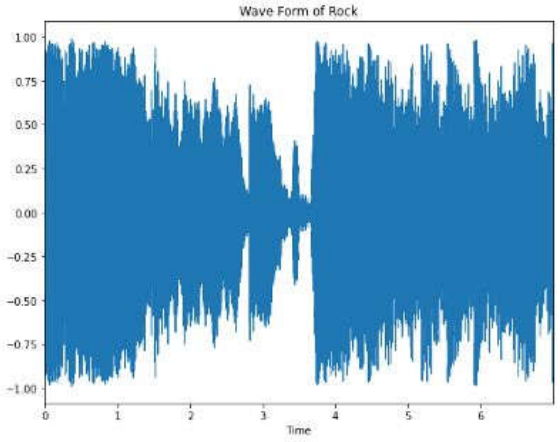
6.3.1 音频的时域特征表示

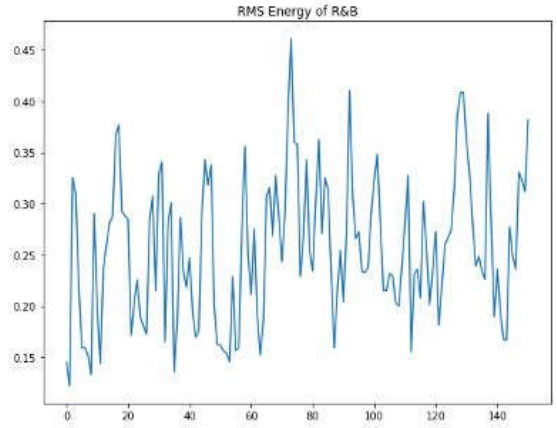
RMS结果

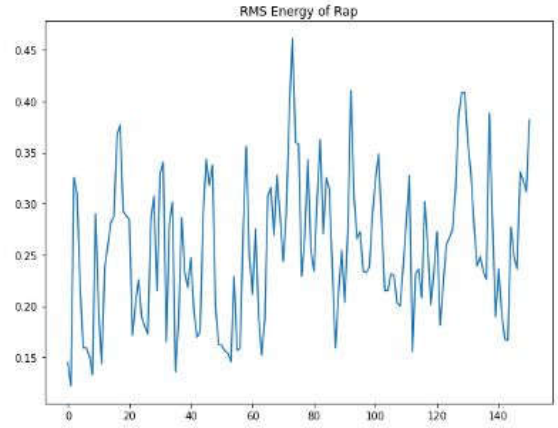
- 均方根波动不像声发射那样剧烈。这个特性使振幅的均方根对异常值更加稳健。

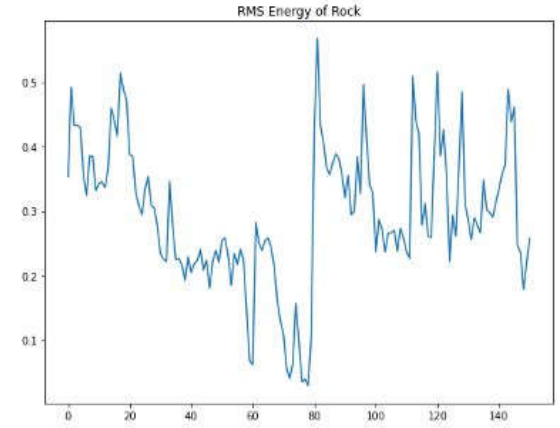












6.3.1 音频的时域特征表示

过零率

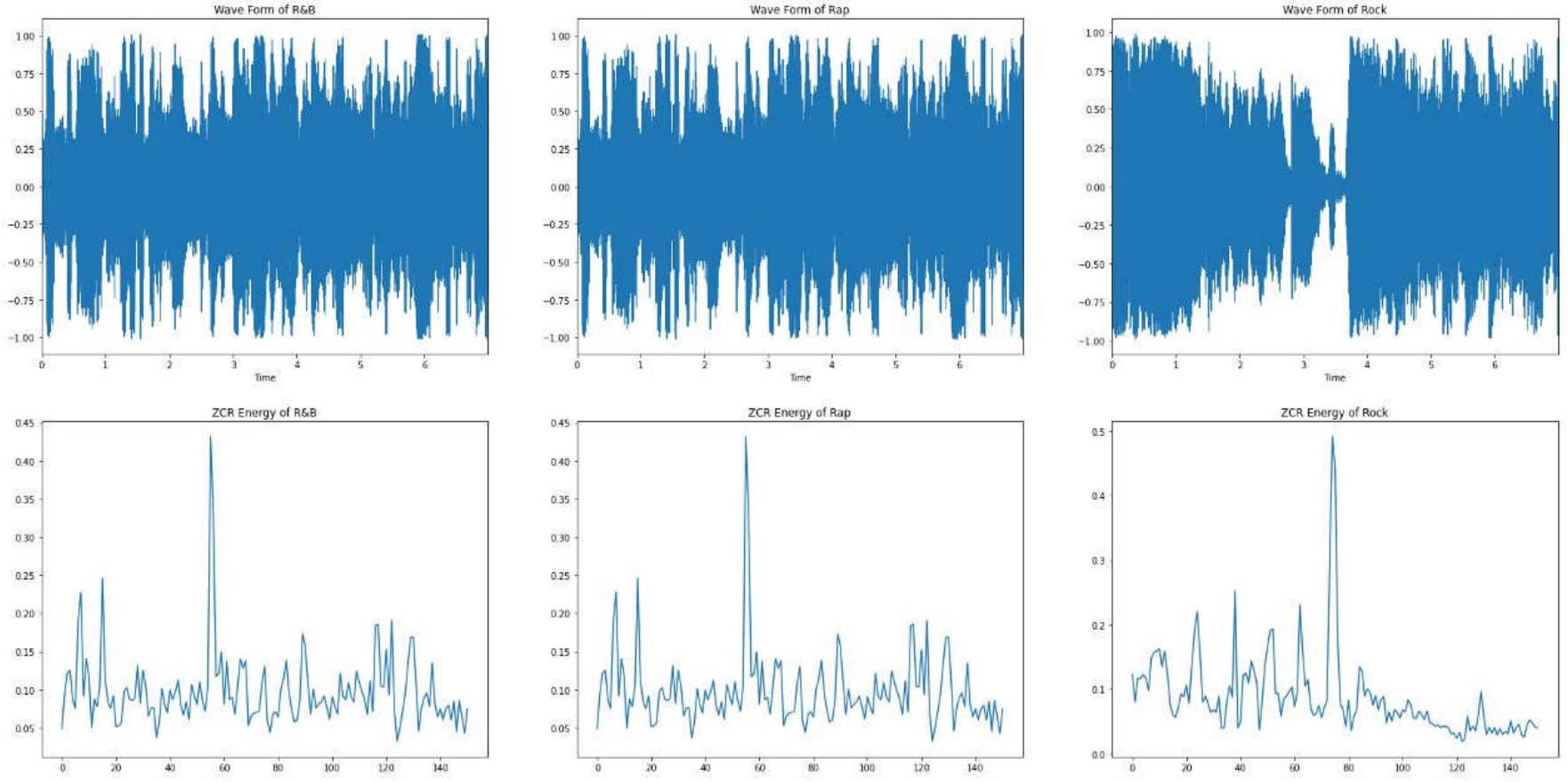
- 过零速率(ZCR)的目的是研究信号的幅值在每一帧中的变化速率。与前两个特性相比，这个特性非常容易提取。

$$ZCR_t = \frac{1}{2}\sum_{k=tk}^{(t+1)k-1} |\operatorname{sgn}(s(k)) - \operatorname{sgn}(s(k+1))|$$

6.3.1 音频的时域特征表示

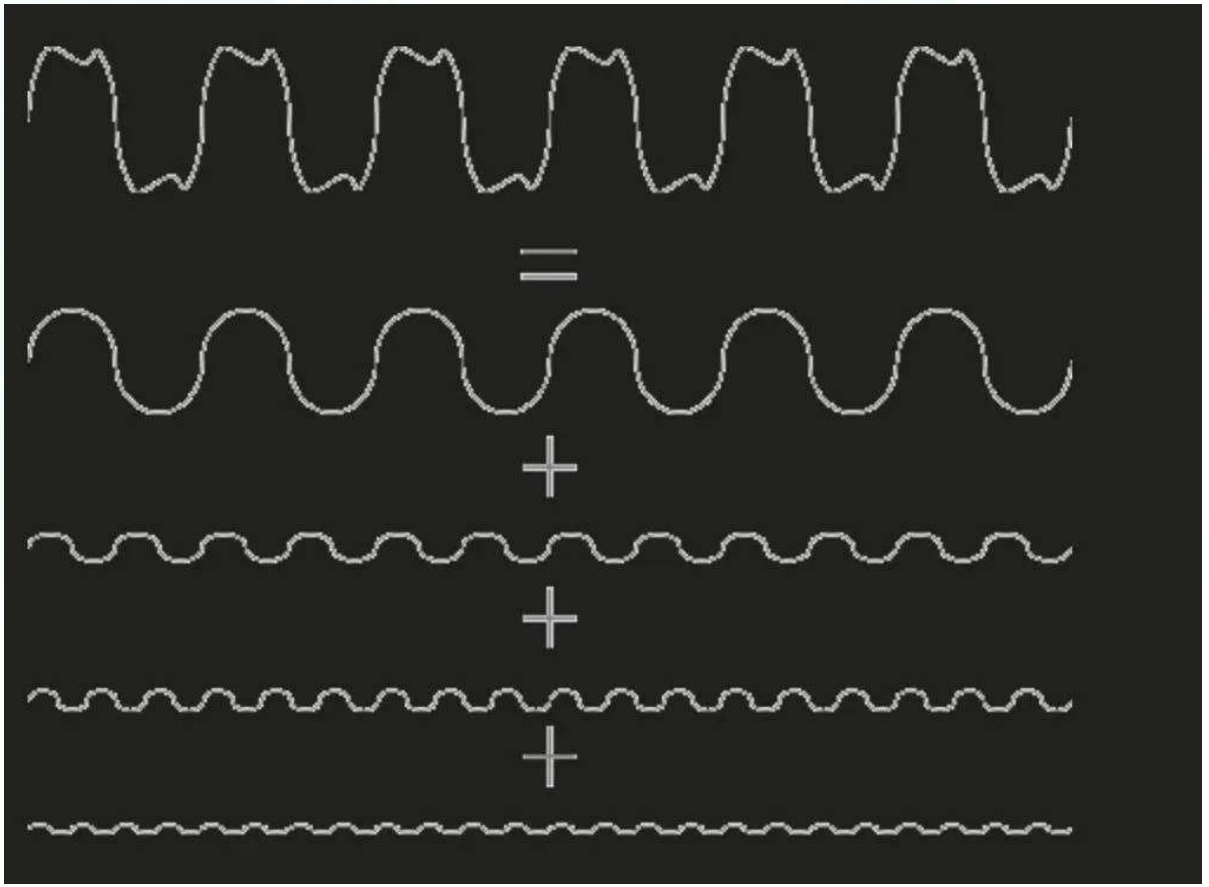
ZCR结果

- ZCR可以很好地检测到这些声音，并且可以检测到音高



6.3.2 音频的频域特征表示

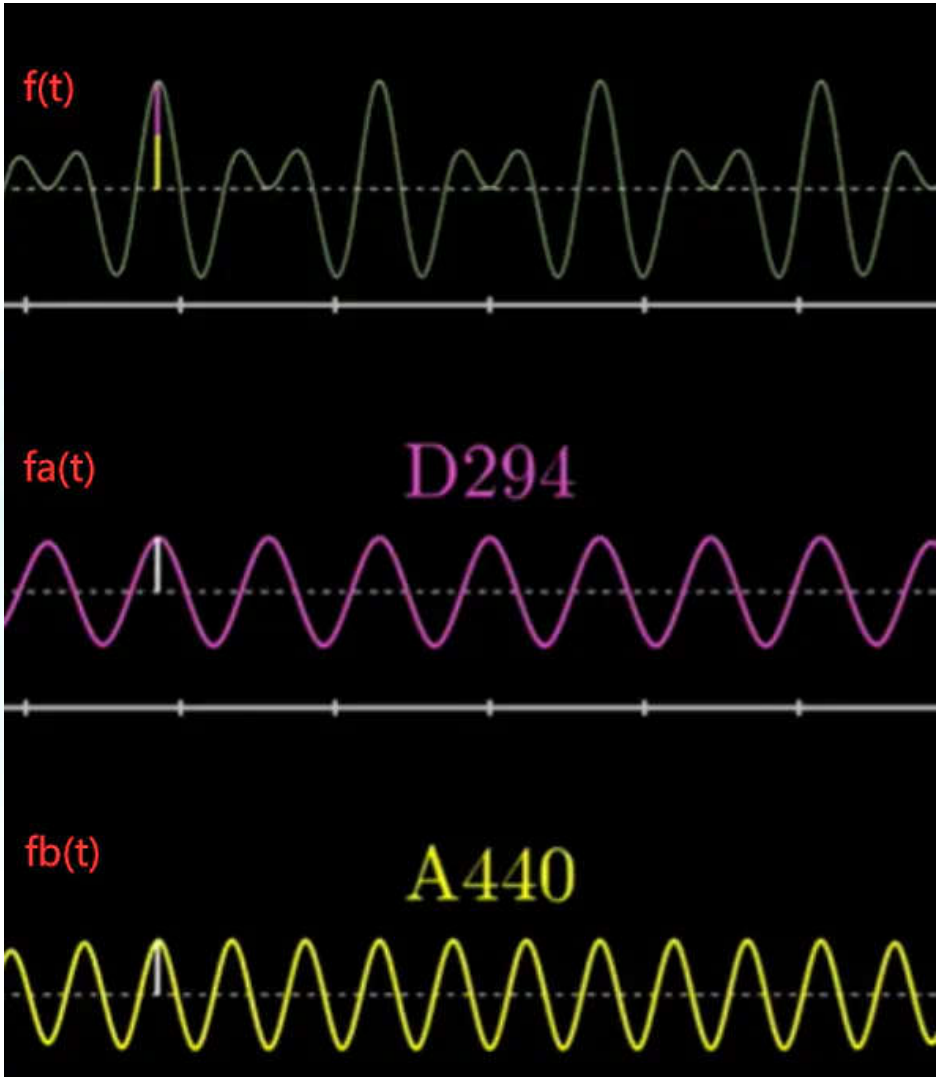
- 指数形式的傅里叶级数公式 $f(t)=\sum_{n=-\infty}^{\infty} F_n e^{jnw_0 t}$,



6.3.2 音频的频域特征表示

音频傅里叶级数例子

- f(t)=fa(t)+fb(t)

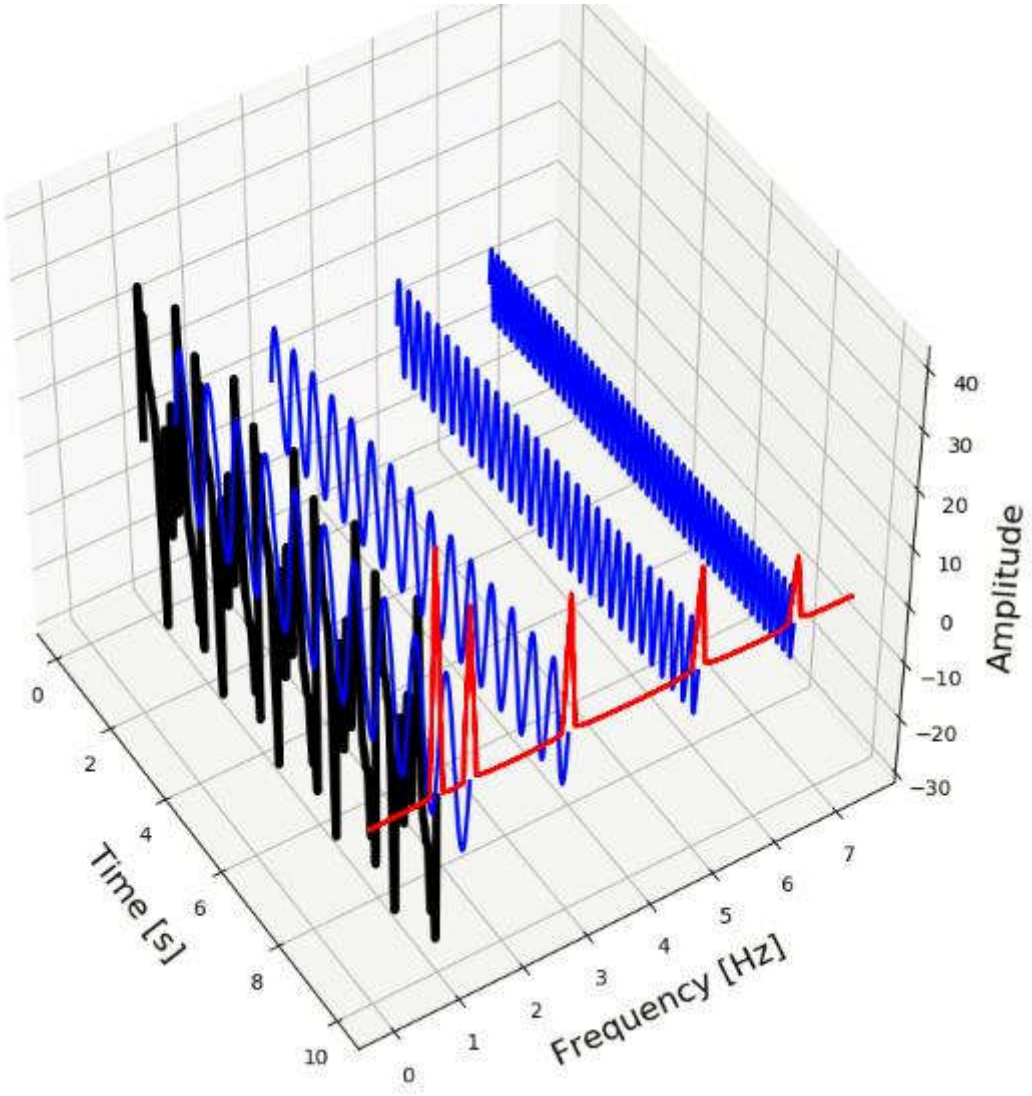


6.3.2 音频的频域特征表示

音频傅里叶变换

$$f(t) = \sum_{k=0}^{N-1} F(k) e^{j2}$$

$$F(k) = \frac{1}{N} \sum_{n=0}^{N-1} f(t)$$



问题3？

- 如何同时获得音频的时域和频域表示？

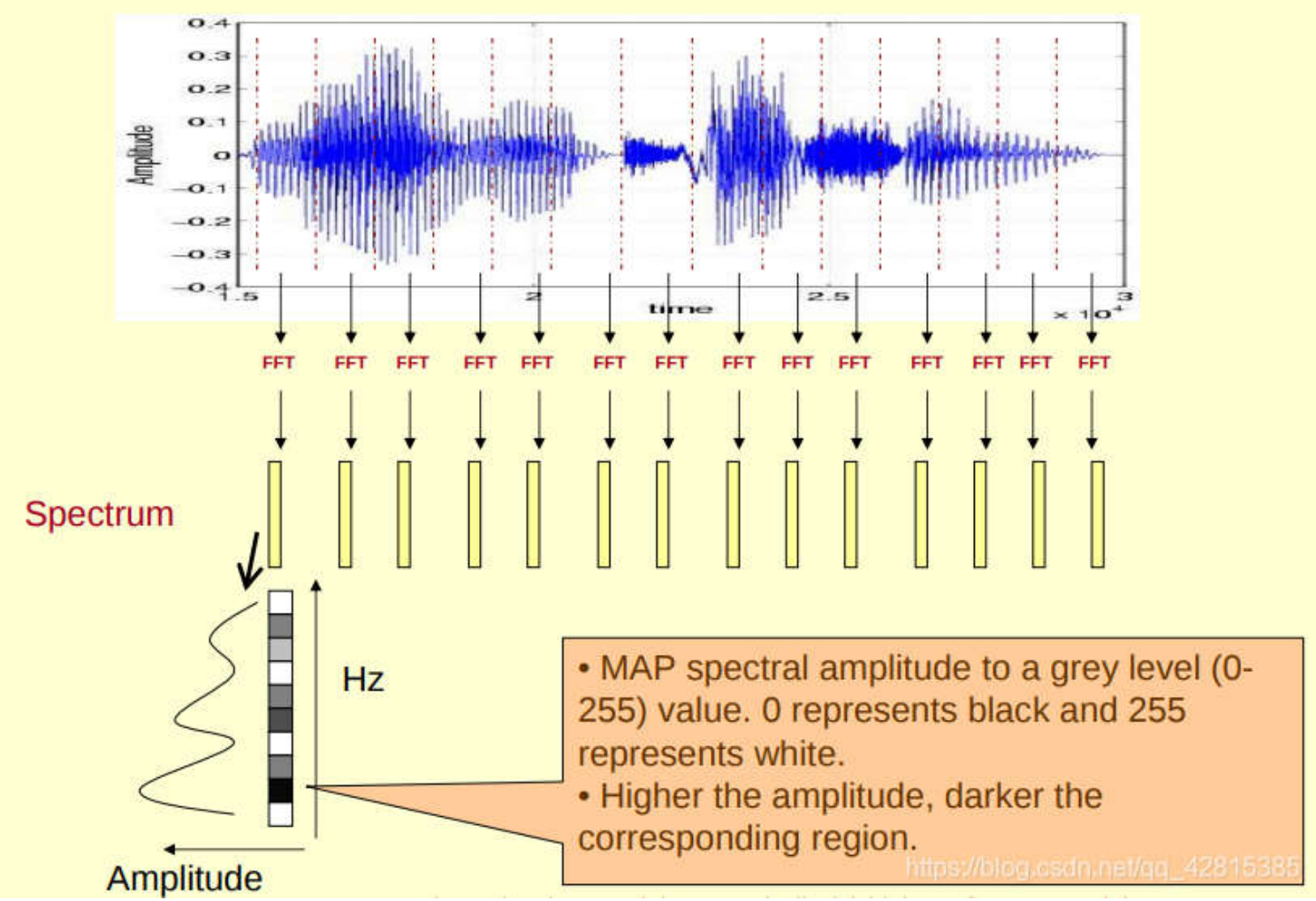
短时傅里叶变换

长时间宽度

短时间宽度

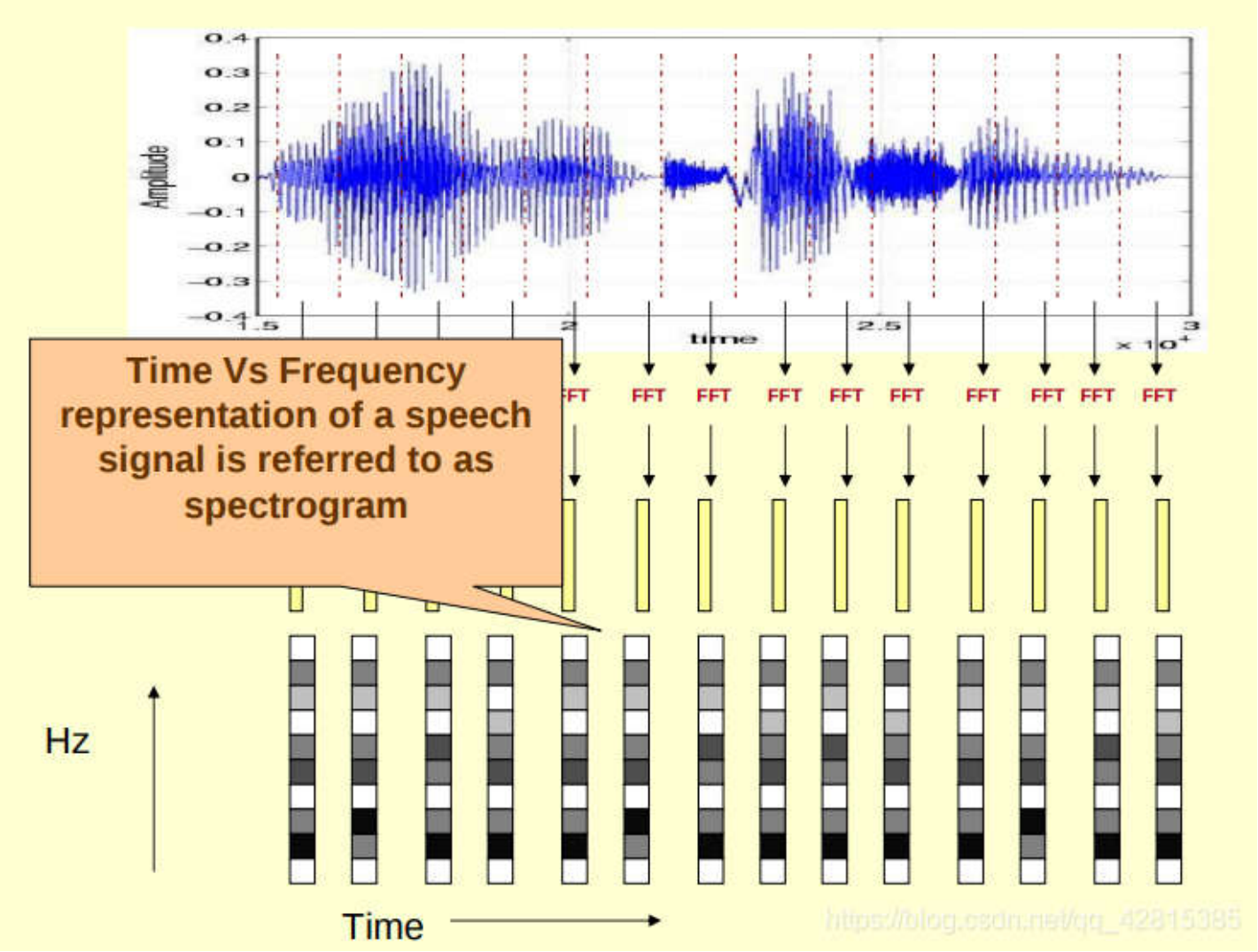
6.3.2 音频的频域特征表示

音频短时傅里叶变换



6.3.2 音频的频域特征表示

音频短时傅里叶变换



6.3.2 音频的频域特征表示

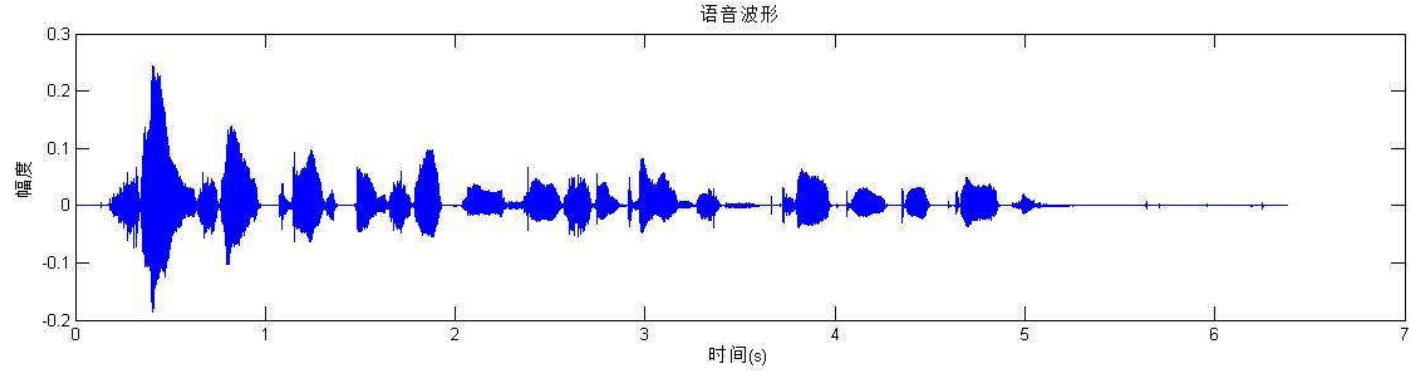
音频时频谱图（语谱图）

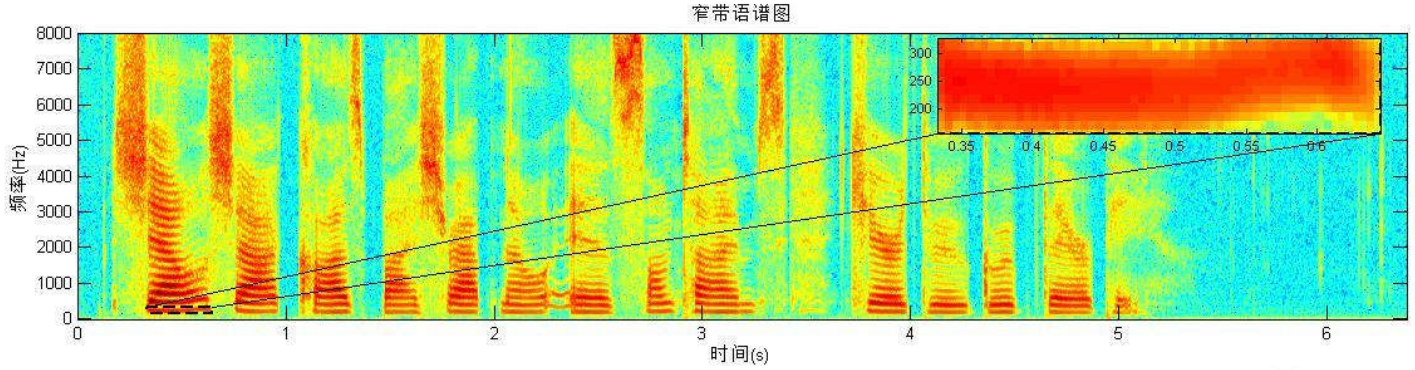
- 首先，什么是语谱图。最通常的，就是语音短时傅里叶变换的幅度画出的2D图。
- 窄带语谱图：带宽小，则时宽大，则短时窗长，窄带语谱图就是长窗条件下画出的语谱图
- 宽带语谱图：“宽带”，正好相反

6.3.2 音频的频域特征表示

窄带语谱图

- 带宽窄，那么在频率上就“分得开”，即能将语音各次谐波“看得很清楚”，即表现为“横线”

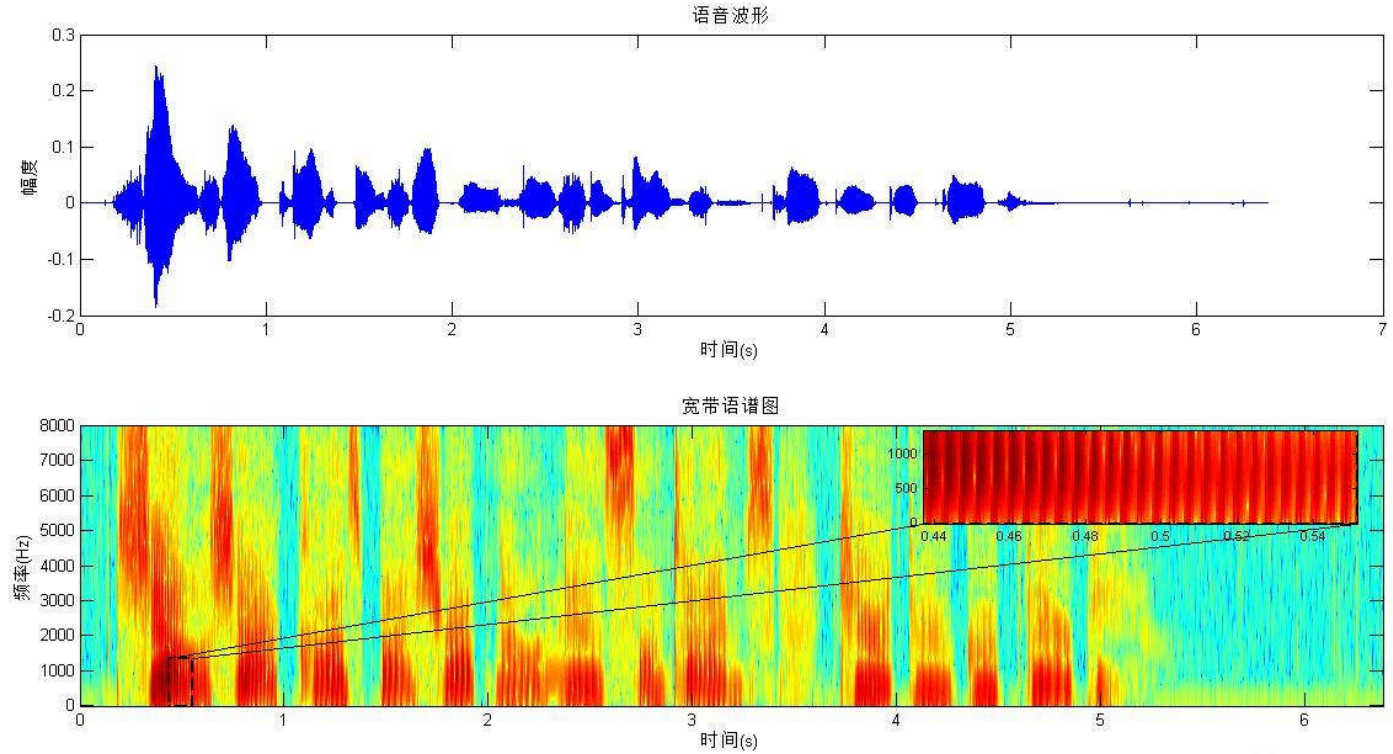




6.3.2 音频的频域特征表示

窄带语谱图

- 时宽窄，那么在时间上就“分得开”，即能将语音在时间上重复的部分“看得很清楚”，即表现为“竖线”

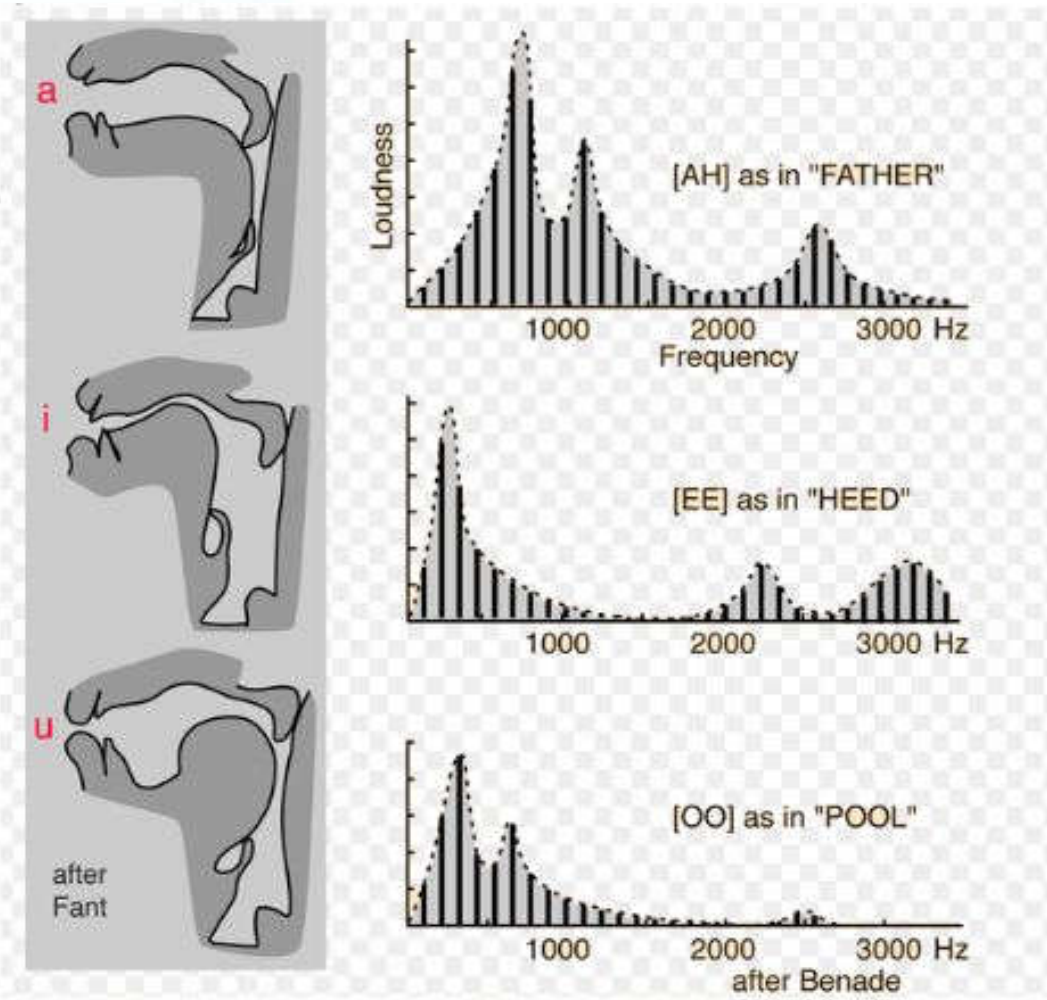


6.3.3 MFCC音频特征

- MFCCs中文名为“**梅尔倒频谱系数**”（Mel Frequency Cepstral Coefficents）是一种在自动语音和说话人识别中广泛使用的特征。它是在1980年由Davis和Mermelstein搞出来的。从那时起。在语音识别领域，MFCCs在人工特征方面影响很大。

共振峰——谐振频率的峰值

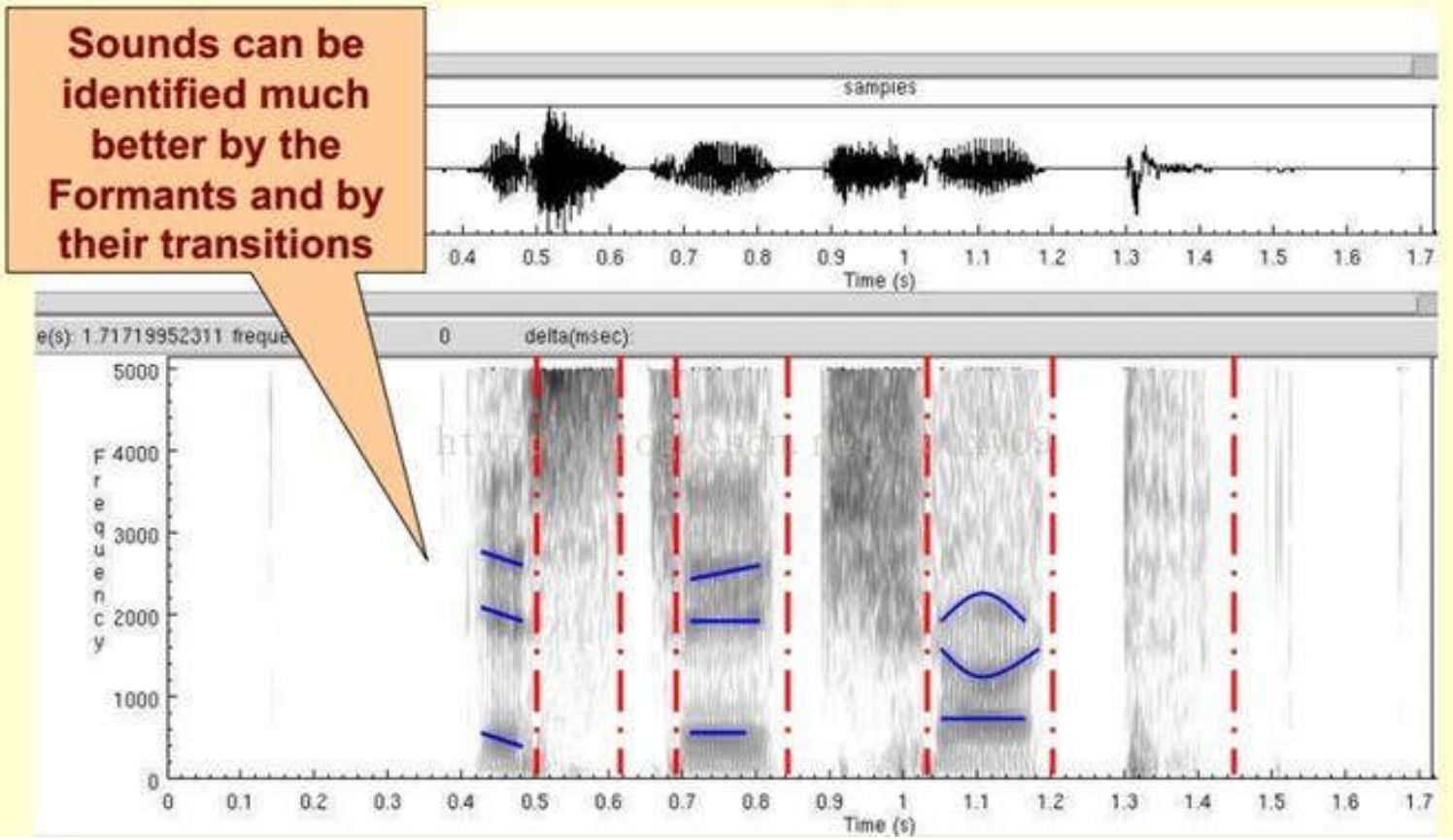
- 当发音的激励频率等于声道的谐振频率时，引起声道共鸣，引起谐振。



6.3.3 MFCC音频特征

基于语谱图的分析

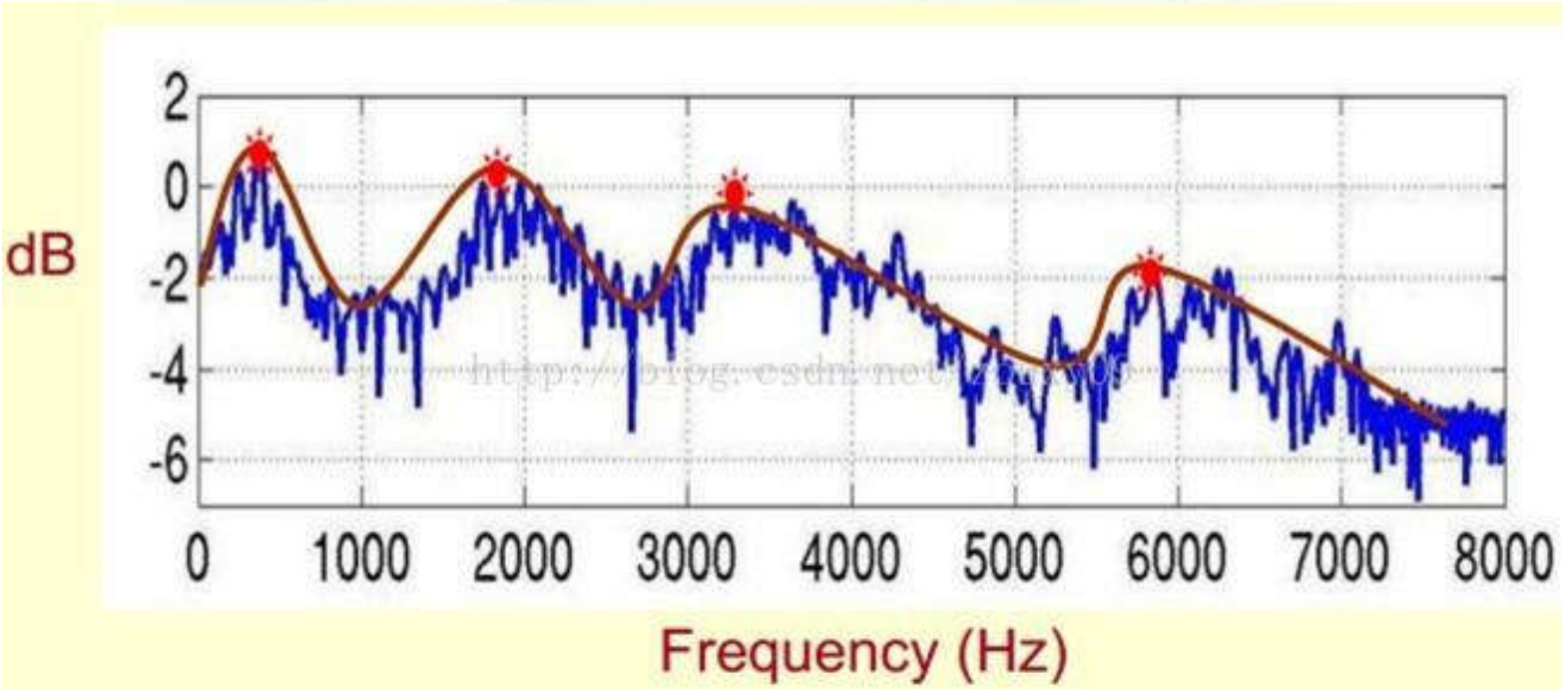
- 下图是一段语音的声谱图，很黑的地方就是频谱图中的峰值（共振峰formants）



6.3.3 MFCC音频特征

提取频谱的包络

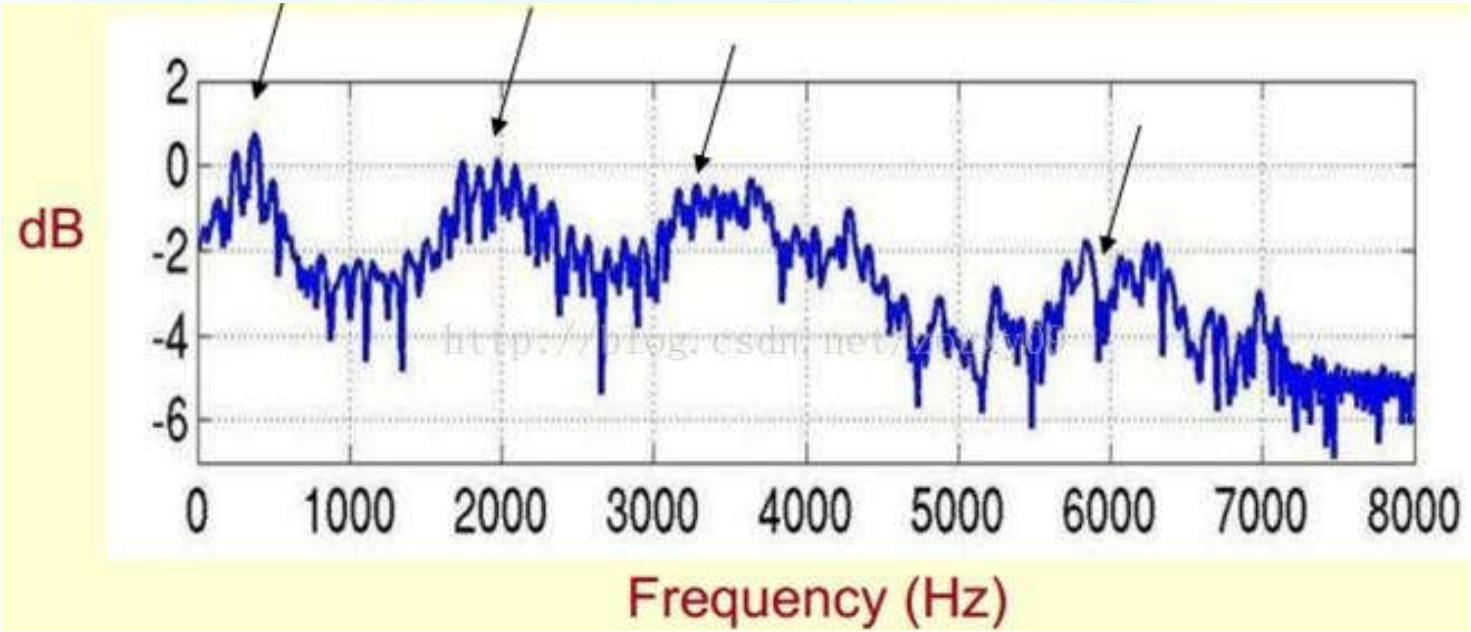
- 提取的不仅仅是共振峰的位置，还得提取它们转变的过程。所以我们提取的是频谱的包络（Spectral Envelope）。这包络就是一条连接这些共振峰点的平滑曲线。



6.3.3 MFCC音频特征

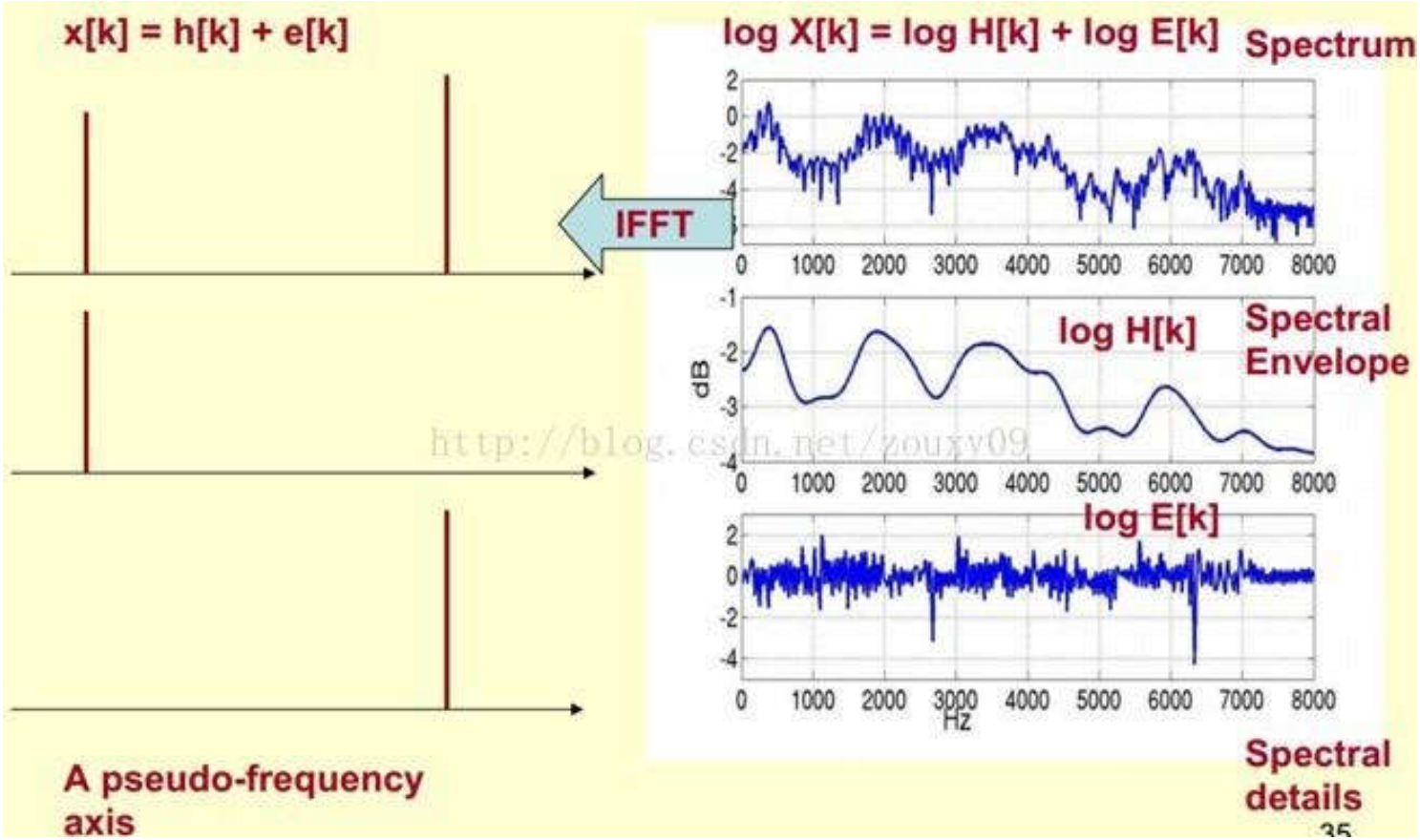
倒谱分析（Cepstrum Analysis）

- 下面是一个语音的频谱图。**峰值就表示语音的主要频率成分，我们把这些峰值称为共振峰**（formants），而**共振峰就是携带了声音的辨识属性**（就是个人身份证一样）。所以它特别重要。用它就可以识别不同的声音。



6.3.3 MFCC音频特征

倒谱分析（Cepstrum Analysis）



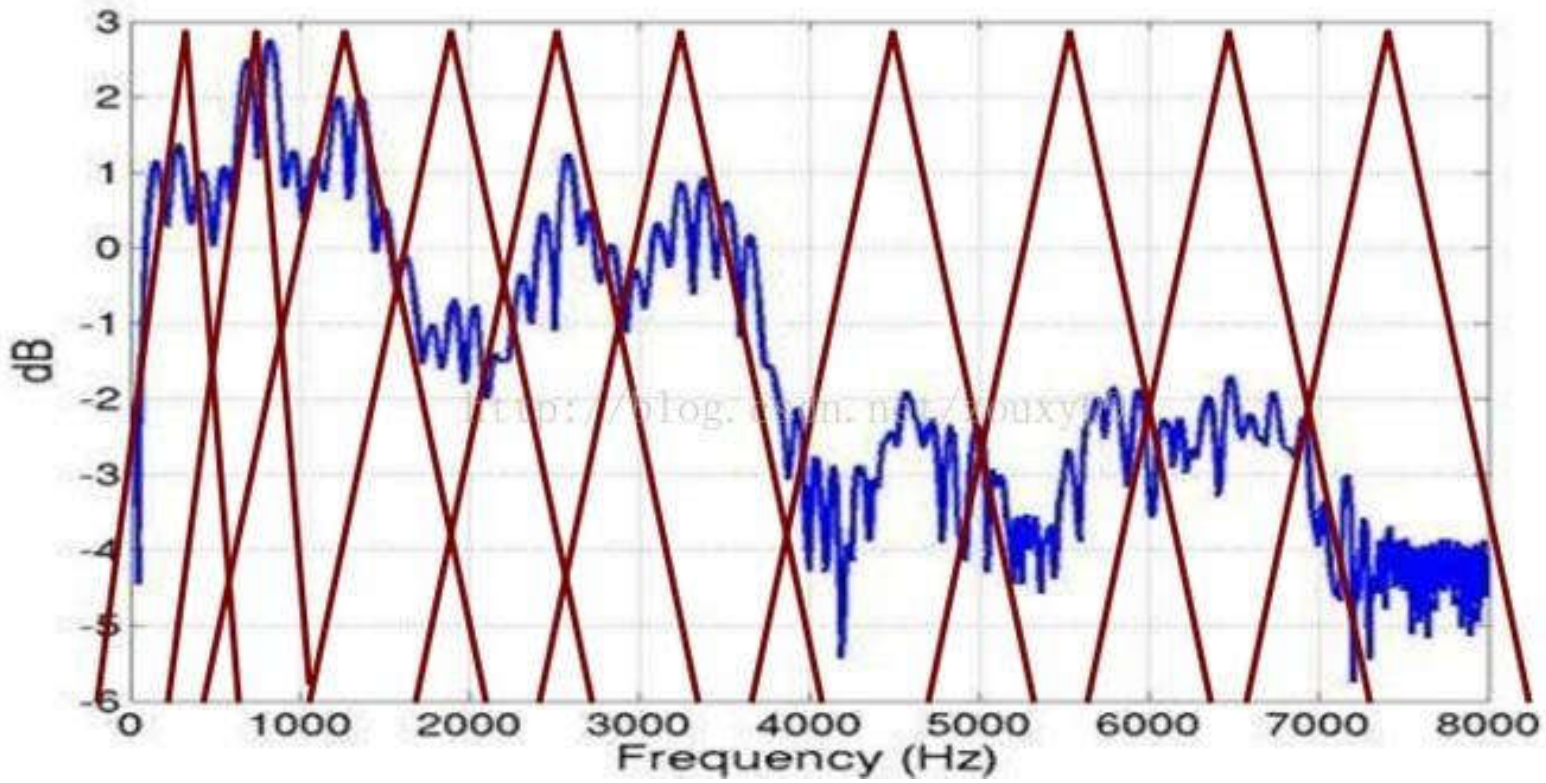
在频谱上做傅里叶变换就相当于逆傅里叶变换Inverse FFT（IFFT）

倒谱分析

6.3.3 MFCC音频特征

Mel频率倒谱系数（Mel-Frequency Cepstral Coefficients）

- Mel频率分析在低频区域有很多的滤波器，他们分布比较密集，但在高频区域，滤波器的数目就变得比较少，分布很稀疏



6.3.3 MFCC音频特征

- 梅尔频率倒谱系数（Mel Frequency Cepstrum Coefficient, MFCC）考虑到了人类的听觉特征，先将线性频谱映射到基于听觉感知的Mel非线性频谱中，然后转换到倒谱上

- 1）先对语音进行预加重、分帧和加窗；
- 2）对每一个短时分析窗，通过FFT得到对应的频谱；
- 3）将上面的频谱通过Mel滤波器组得到Mel频谱；
- 4）在Mel频谱上面进行倒谱分析（取对数，做逆变换，实际逆变换一般是通过DCT离散余弦变换来实现，取DCT后的第2个到第13个系数作为MFCC系数），获得Mel频率倒谱系数MFCC，这个MFCC就是这帧语音的特征；

6.3.3 MFCC音频特征