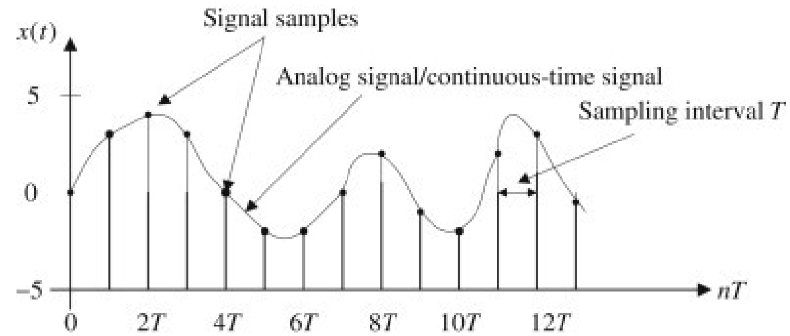
Introduction to Speech Signal

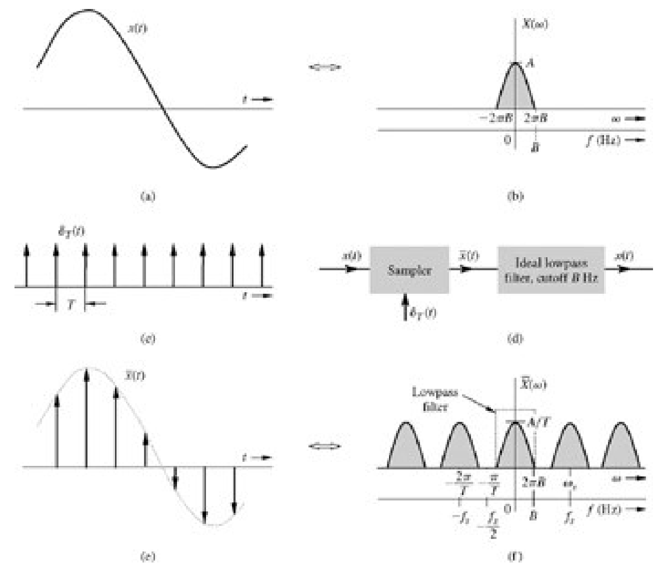
Meng Yu

08/29/2022

Speech Signal Representation

- Represent continuous signal into discrete form
- Max measurable frequency is half sampling rate (Shannon sampling theorem)
- Quantization Representing real value of each amplitude as integer 8-bit (-128 to 127) or 16-bit (-32768 to 32767)

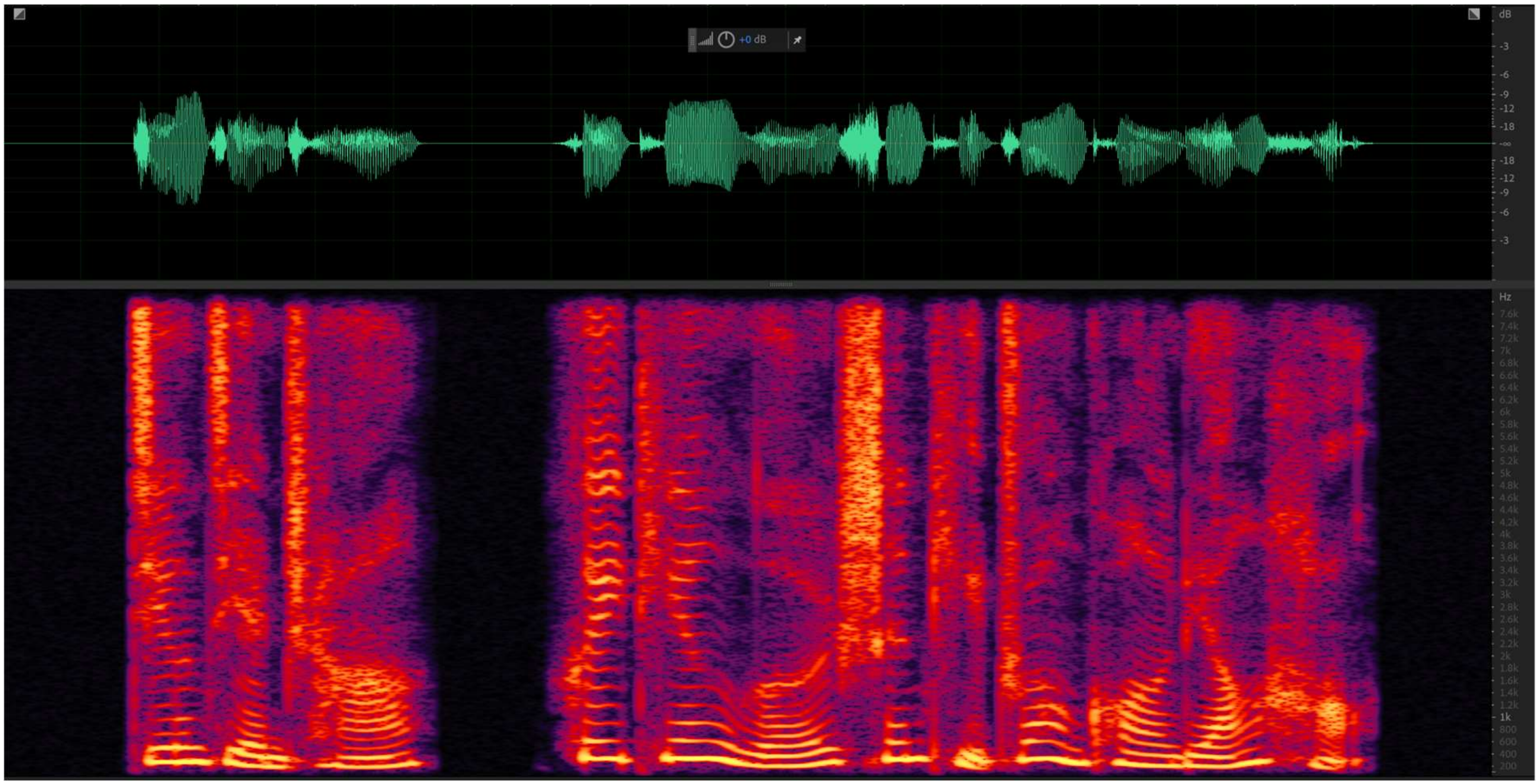




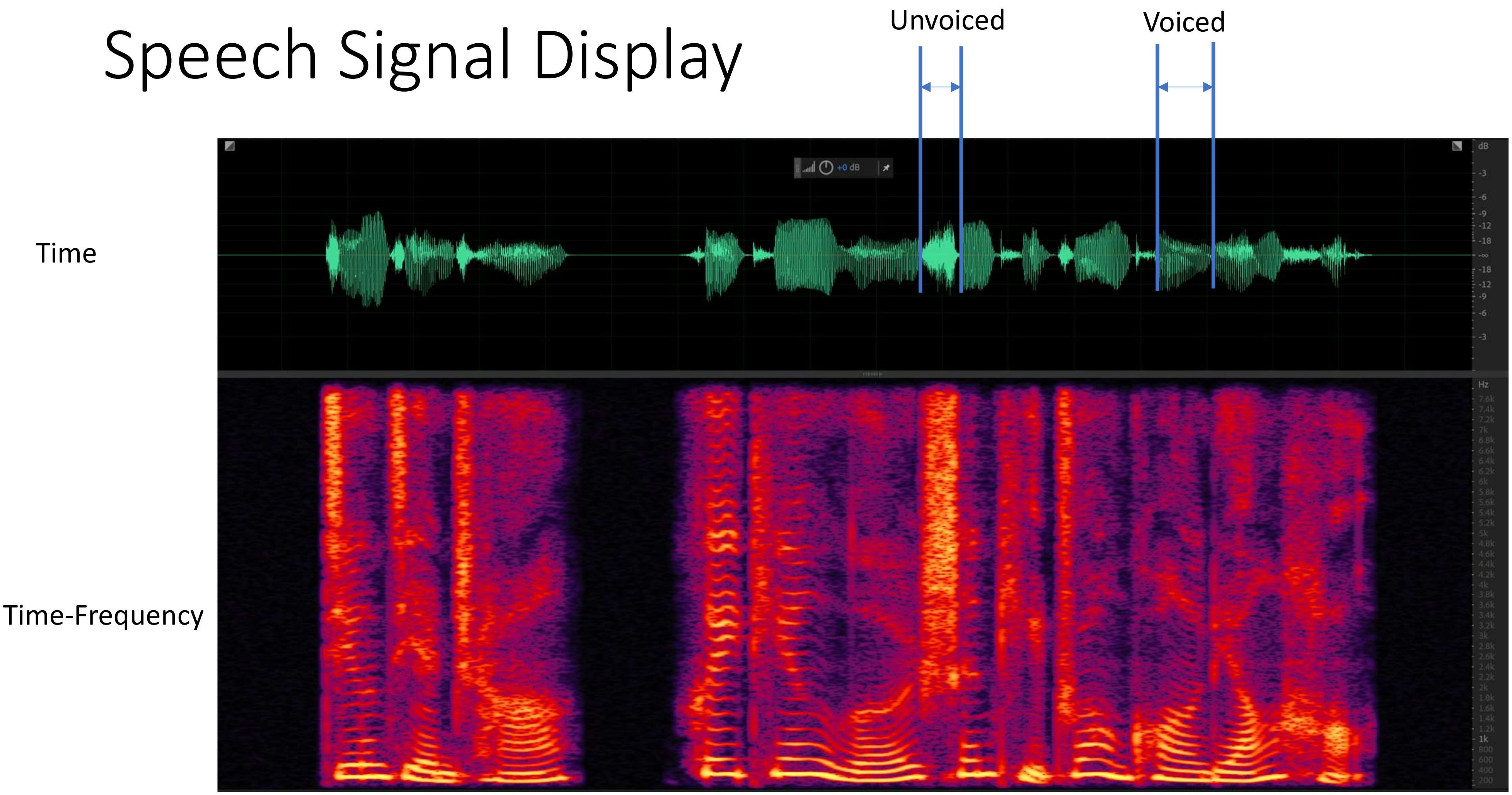
Speech Signal Display

Time

Time-Frequency



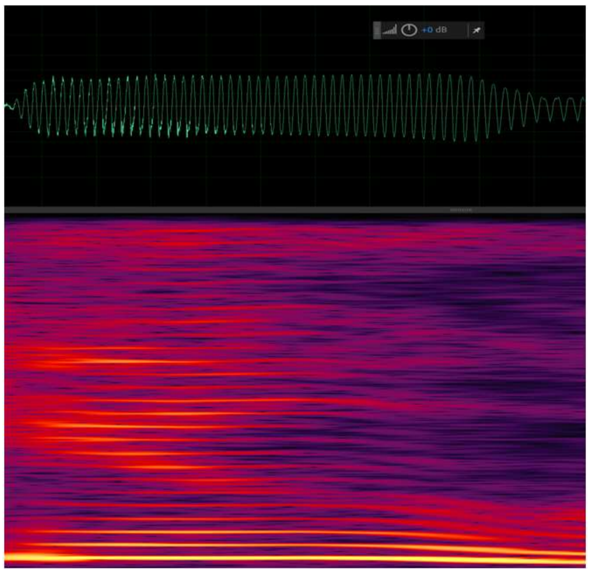
Speech Signal Display



Voiced

Sinusoid in time domain

Harmonics in frequency domain



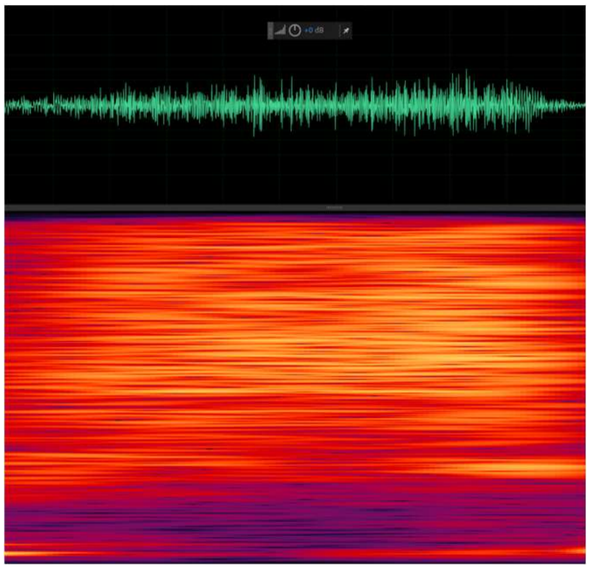
Time and Time-Frequency domain

Frequency response

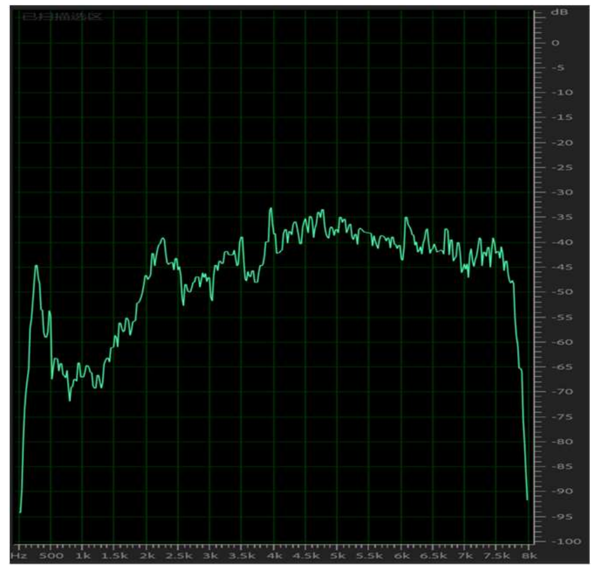
浊音声带紧绷，气流来了后，张弛振动，周期性的开启和闭合，形成准周期性的脉冲状空气流（周期为基音周期）。声带越短、厚度越薄、张力越大，则音调越高，即浊音的基音频率越高。男性基音频率50—250Hz，女性基音频率100—500Hz

Unvoiced

Like a white noise



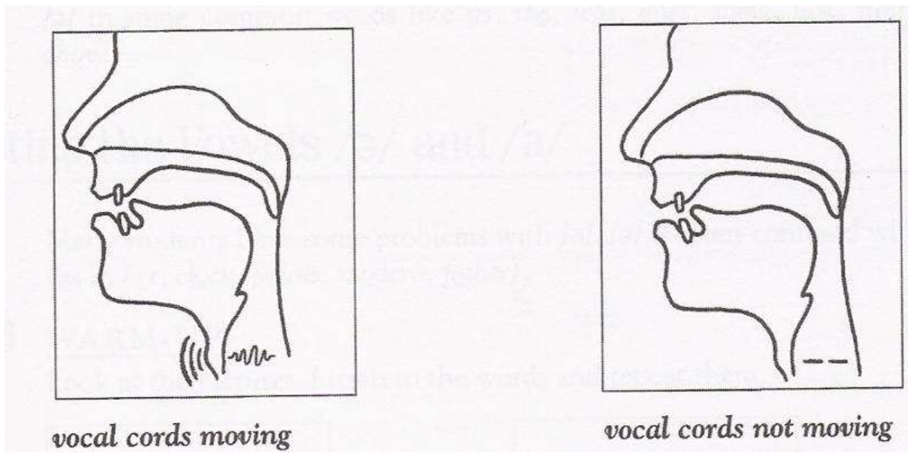
Time and Time-Frequency domain

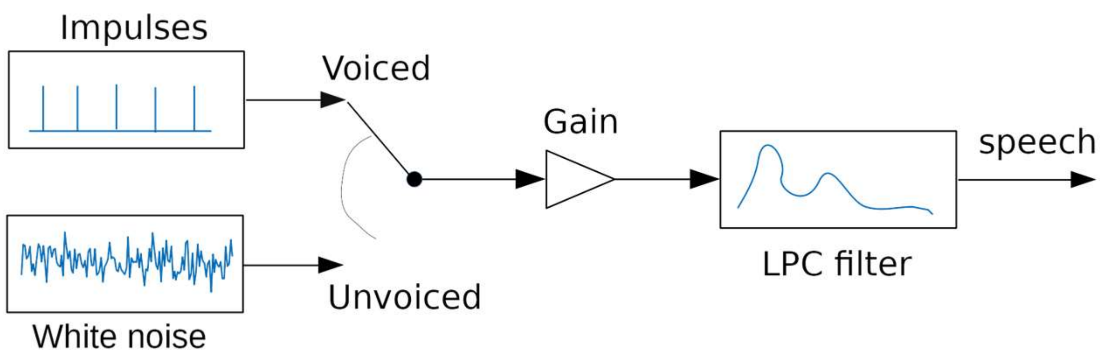


Frequency response

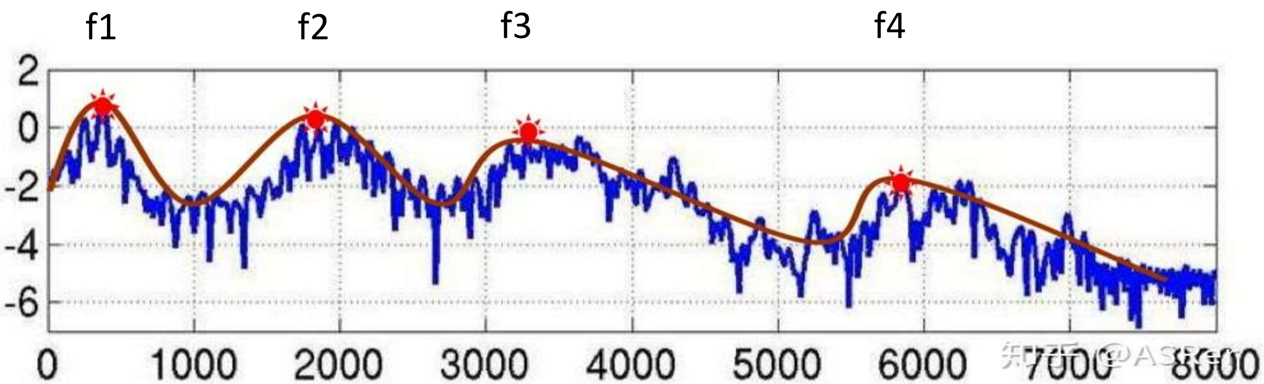
清音声带完全舒展，气流来了后，两种情况：第一，声道某个部位收缩为狭窄通道，声流高速冲过，形成摩擦音或者清音；第二，声道某部位完全闭合（如闭嘴），则形成爆破音。

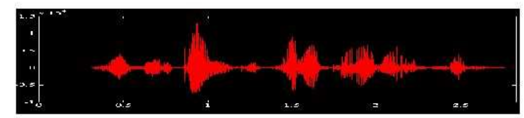
Speech Production

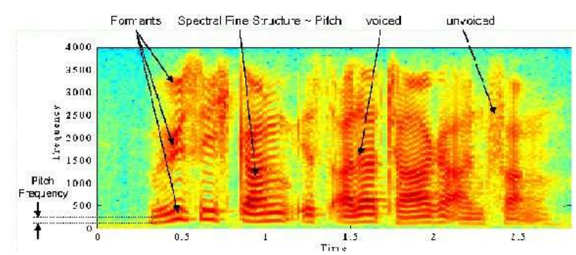




Formant (共振峰)



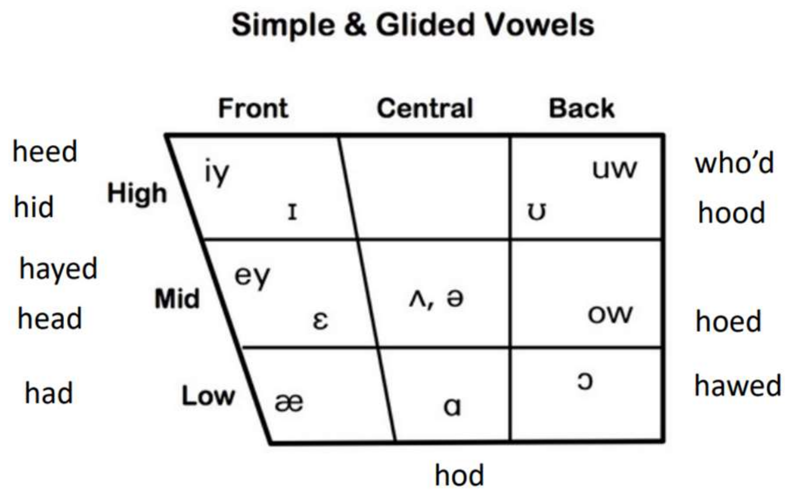


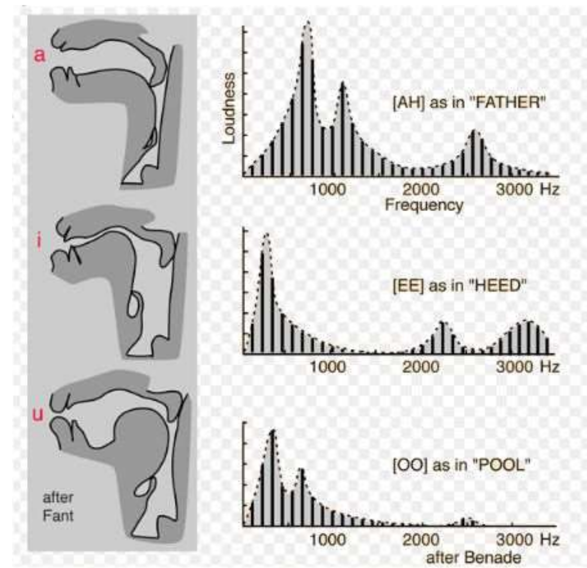


通常有3-5个共振峰
开口度越大，f1越高；
舌位越靠前，f2越高；
不圆唇元音的f3比圆唇元音高。

Vowel

- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/
- Tense/lax:
  - Tense: e.g., /i, e, u, o, ɔ, ɑ/
  - Lax: e.g., /ɪ, ɛ, æ, ə/

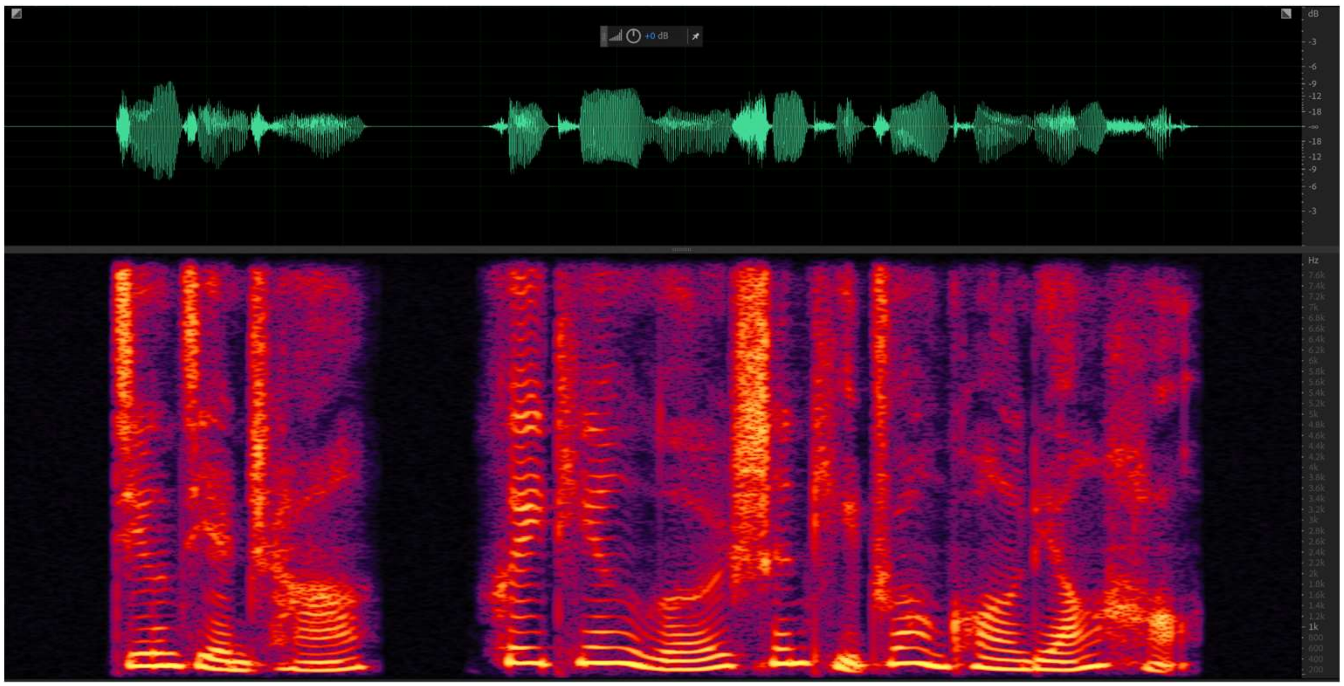




Consonant

- 辅音也叫“子音”。发音时，气流从肺中呼出，经过声门、咽腔、口腔或鼻腔时，受到各个器官不同程度的阻碍，不能畅通。由于阻碍部位（发音部位）和阻碍的方法及除去阻碍的方式（发音方法）不同，造成不同的辅音
- – 塞音 Stops: /p, t, k, b, d, g/
- – 擦音 Fricatives: /f, s, v, z/
- – 塞擦音 Affricates: /ts, dz/
- – 近音/边音 Approximants/Liquids: /l, r, w, j/
- – 鼻音 Nasals: /m, n, ng/

Consonant contains voiced sound

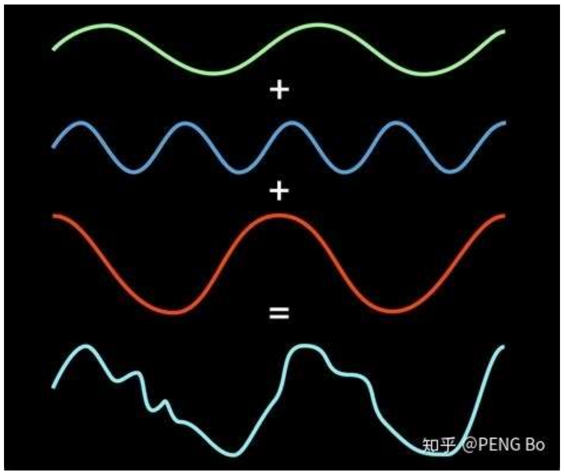


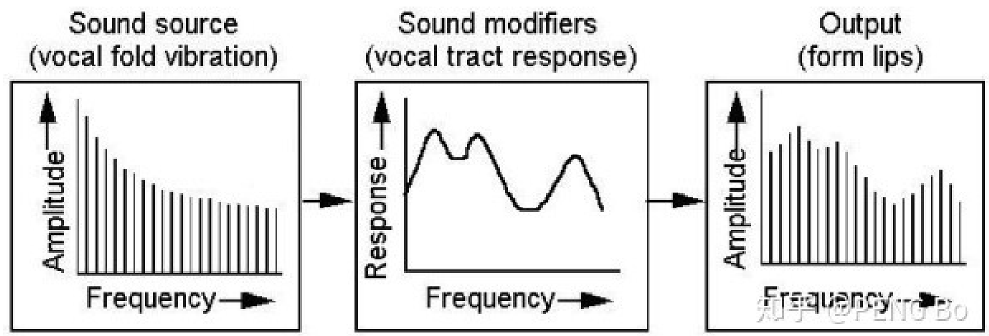
j ch h sh k l h

浊音会在清音的基础上有周期性的精细结构

Voiced Vibration		**Voiceless** No Vibration	
b	→ bat	p	→ pat
d	→ dot	t	→ tall
g	→ gap	k	→ cap
v	→ vine	f	→ fine
th	→ this	th	→ thin
z	→ zoo	s	→ sue
j	→ gym	sh	→ shore
m	→ mail	h	→ hot
n	→ nail	ch	→ chip
ng	→ sing		
l	→ let		
r	→ root		
w	→ wet		
y	→ yard		

Speech Signal

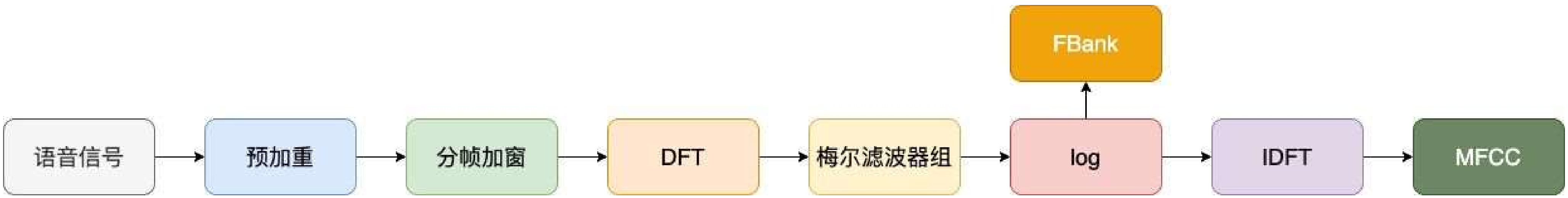




Green Function

- 对线性算子 L， 在点源 δ 作用下的输出（或响应）就是格林函数G，即： LG=δ
- 声波波动问题，线性算子为 $L = \frac{\partial^2}{\partial t^2} - c^2 \nabla^2$
- 若已知格林函数与源分布（包括时间上与空间上）， 则可通过格林函数与源的卷积求得在此源作用下系统的输出（或响应）
- Lϕ=Q ,其中 L 是线性算子， Q 为源分布， ϕ 为待求输出。利用卷积的性质，可得： ϕ=ϕ∗δ=ϕ∗(LG)=(Lϕ)∗G=Q∗G .

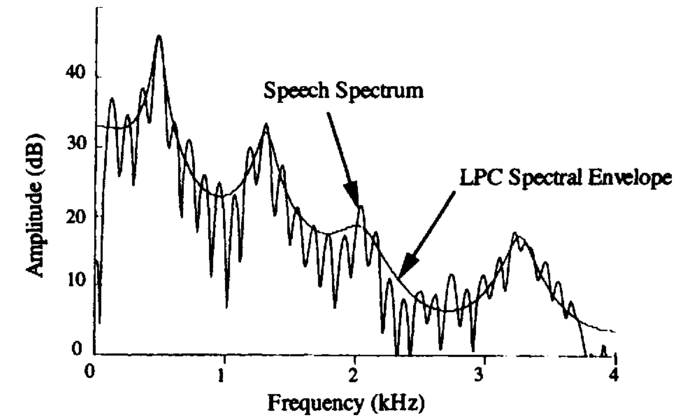
Feature Extraction



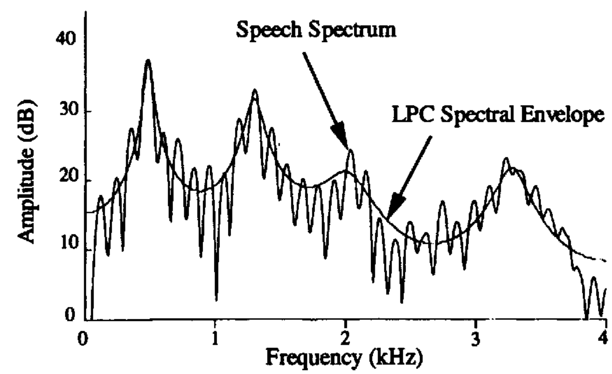
Pre-emphasis

- Boost energy in the high frequencies
- Spectrum for voiced segments has more energy at low frequencies than high frequencies, called spectral tilt, caused by glottal pulse

$$y[n] = x[n] - \alpha x[n-1], \qquad 0.9 \leq \alpha \leq 1.0$$

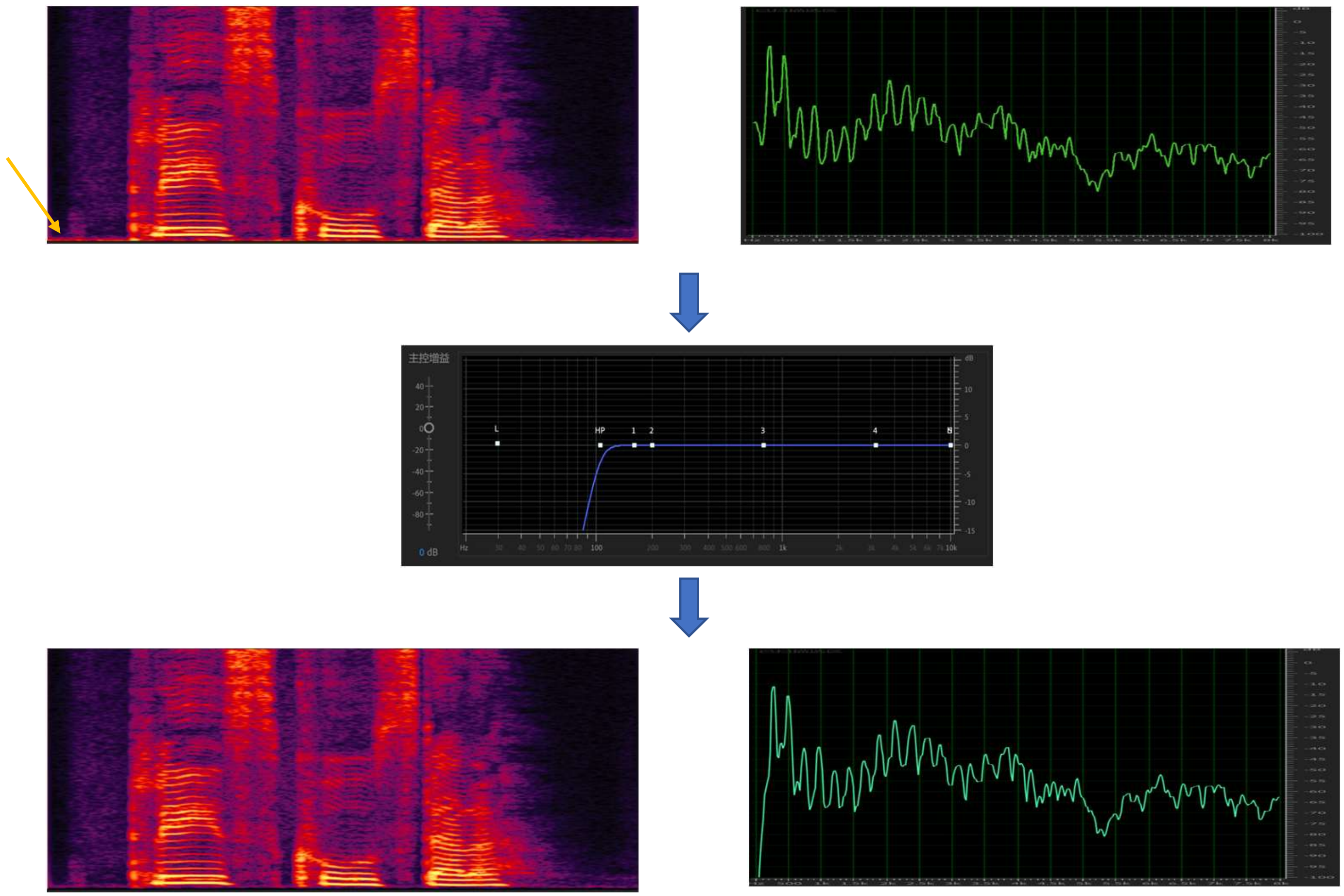


(a) no pre-emphasis



(b) with pre-emphasis

High Pass Filter



Frame Segment

- Why divide speech signal into successive overlapping frames?

  Speech is not a stationary signal

  We want information about a small enough region that the spectral information is a useful cue.

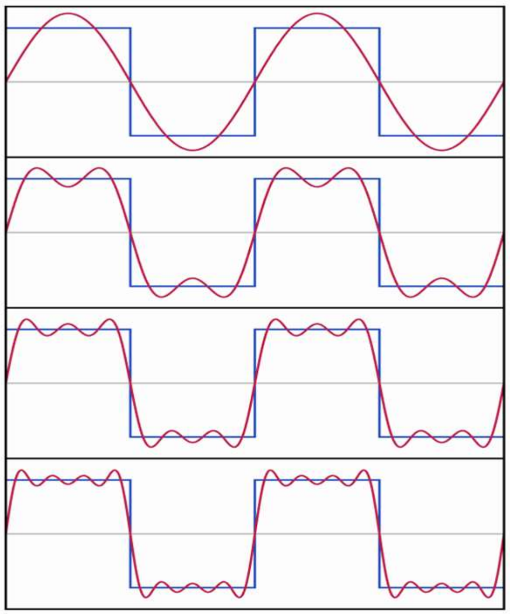
- Frames

  Frame size: typically, 10 -25 ms

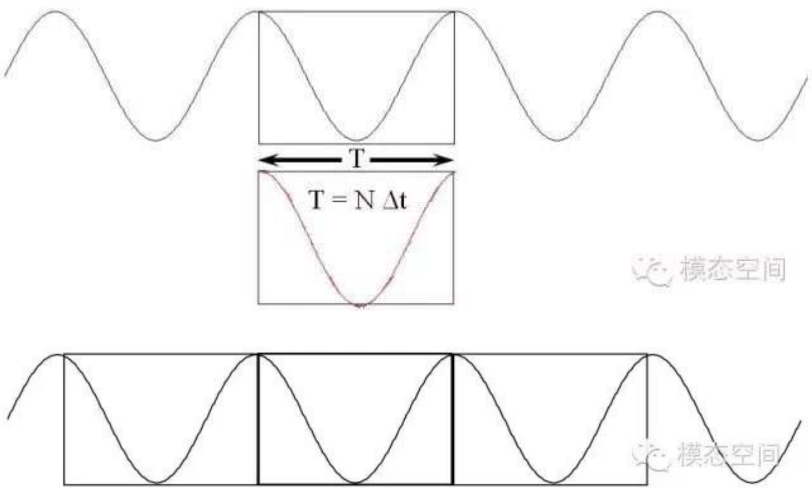
  Frame shift: the length of time between successive frames, typically, 5 -10 ms

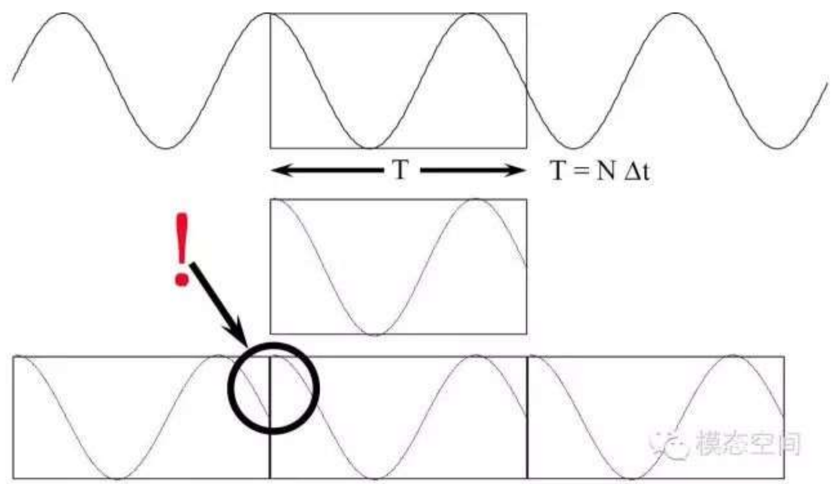
Fourier Transform

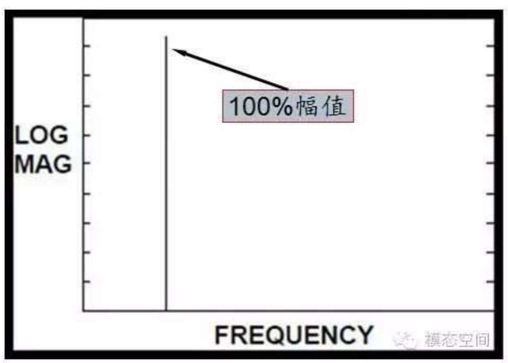
- 周期信号，表示为可数个正弦波的叠加
- 对于非周期信号，我们不能简单地将它展开为可数个正弦波的叠加，但是可以利用傅里叶变换展开为不可数的正弦波的叠加



Fourier Transform on Truncated Signal







假设原始信号的频率为***f*** Hz，则周期为**1/*f*** s。因为截取的时间长度***T***为信号周期的整数倍（假设为k倍），即

$$T=k/f$$

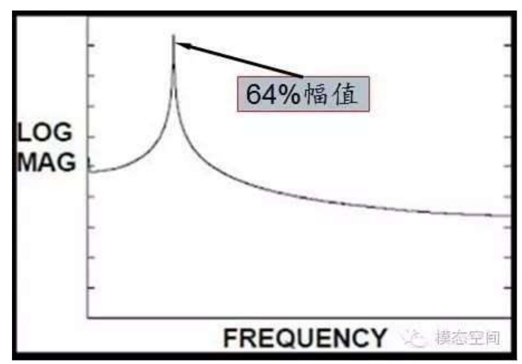
而频率分辨率为**1/*T***，即

$$\Delta f=1/T=f/k$$

因而，信号的频率成分

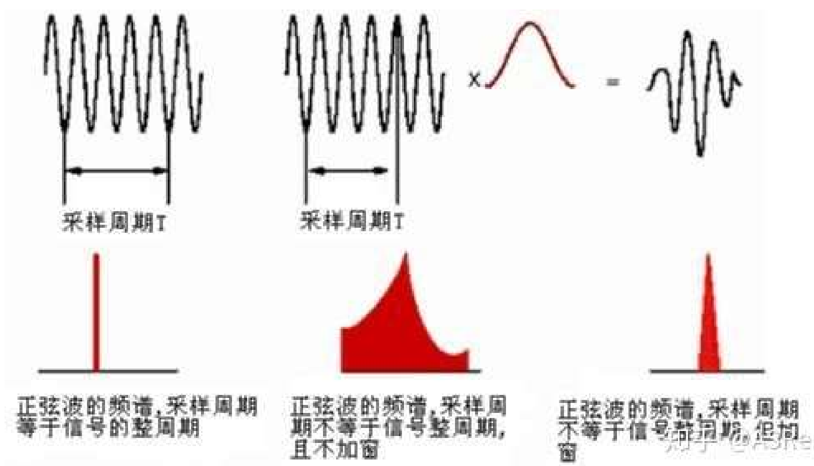
$$f=k*\Delta f$$

即信号的频率成分为频率分辨率**Δ*f***的整数倍，也就是说频谱图中有一条谱线与信号的频率成分相同，这也就是所谓的信号“压谱线”



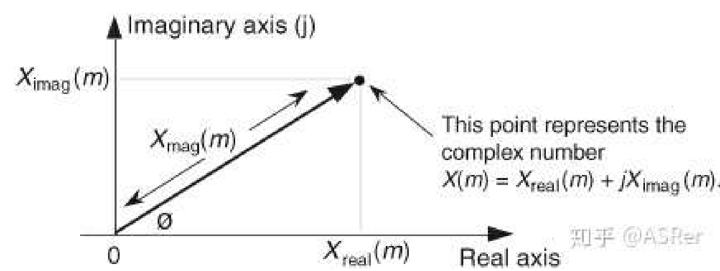
Window for Fourier Transform

- 避免频谱泄露
- 频谱泄露就是分析结果中，出现了本来没有的频率分量。比如说，50Hz 的纯正弦波，本来只有一种频率分量，分析结果却包含了与50Hz频率相近的其它频率分量。



Discrete Fourier Transform

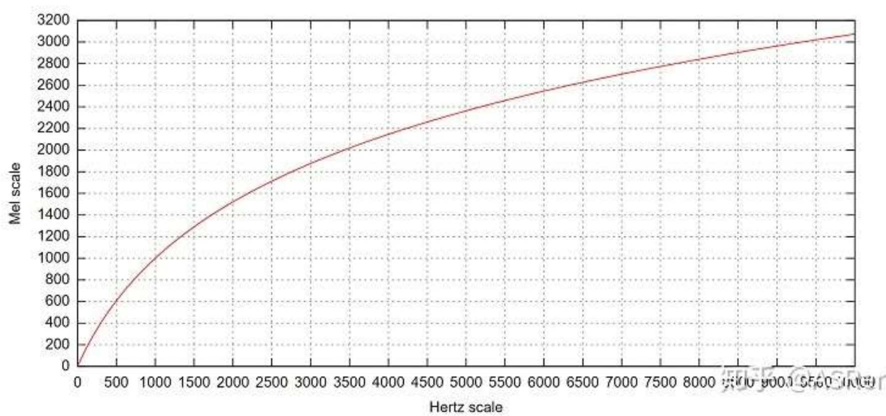
- Input:
  - Windowed signal x[n]…x[m]
- Output:
  - For each of N discrete frequency bands
  - A complex number X[k] representing magnitude and phase of that frequency component in the original signal Discrete Fourier Transform (DFT)
- DFT: $X(m) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nm/N}$

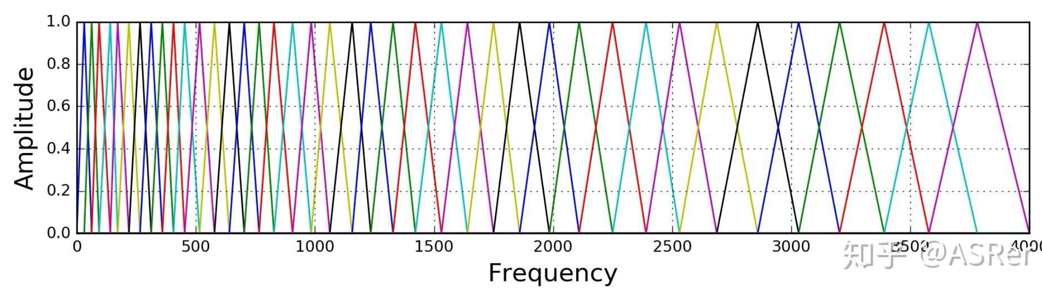


Mel Scale

- Mel-scale Human hearing is not equally sensitive to all frequency bands
- Less sensitive at higher frequencies, roughly > 1000 Hz
- Human perception of frequency is non-linear
- 在 Mel 频域内，人的感知能力为线性关系，如果两段语音的 Mel 频率差两倍，则人在感知上也差两倍
- Mel 频率与频率 Hz 转换的公式如下：

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





40dim Mel Filter Bank

Reference

- https://zhuanlan.zhihu.com/p/147386972
- https://www.semanticscholar.org/paper/Noise-Cancellation-for-CELP-Voice-Encoders-in-an-Heide/f03c7d2d677f6ee13bec09e435b9cc694885a8ca
- https://slidetodoc.com/lsa-352-speech-recognition-and-synthesis-dan-jurafsky-5/
- https://www.sciencedirect.com/topics/engineering/original-signal-spectrum
- https://mp.weixin.qq.com/s/aLSmlrgQF7FBxh_YXXfq6w
- https://zhuanlan.zhihu.com/p/40329331
- https://blog.csdn.net/weixin_42846157/article/details/104486434
- https://jmvalin.ca/demo/lpcnet/
- https://courses.engr.illinois.edu/ece417/fa2017/ece417fa2017lecture8.pdf
- http://campusweb.howardcc.edu/ehicks/YE618/Mastering%20American%20Pronunciation/Voiced_Voiceless_Sounds/Voiced_Voiceless_Sounds_print.html
- https://myenglishfaves.blogspot.com/2017/05/voiceless-and-voiced-consonants-chart.ht
- https://www.zhihu.com/question/24190826