浙江大学 计算机学院

数字媒体与网络技术

数字语音处理III

浙江大学计算机学院

群名称：数字视音频处理-数2019媒
群　号：398550295

杨莹春

yyc@zju.edu.cn, QQ:1169244241

QQ群：数字视音频处理-数2019媒（398550295）

验证信息/群名片：姓名学号口音

浙江大学外经贸楼520

2019年9月25日

数字语音处理课程安排

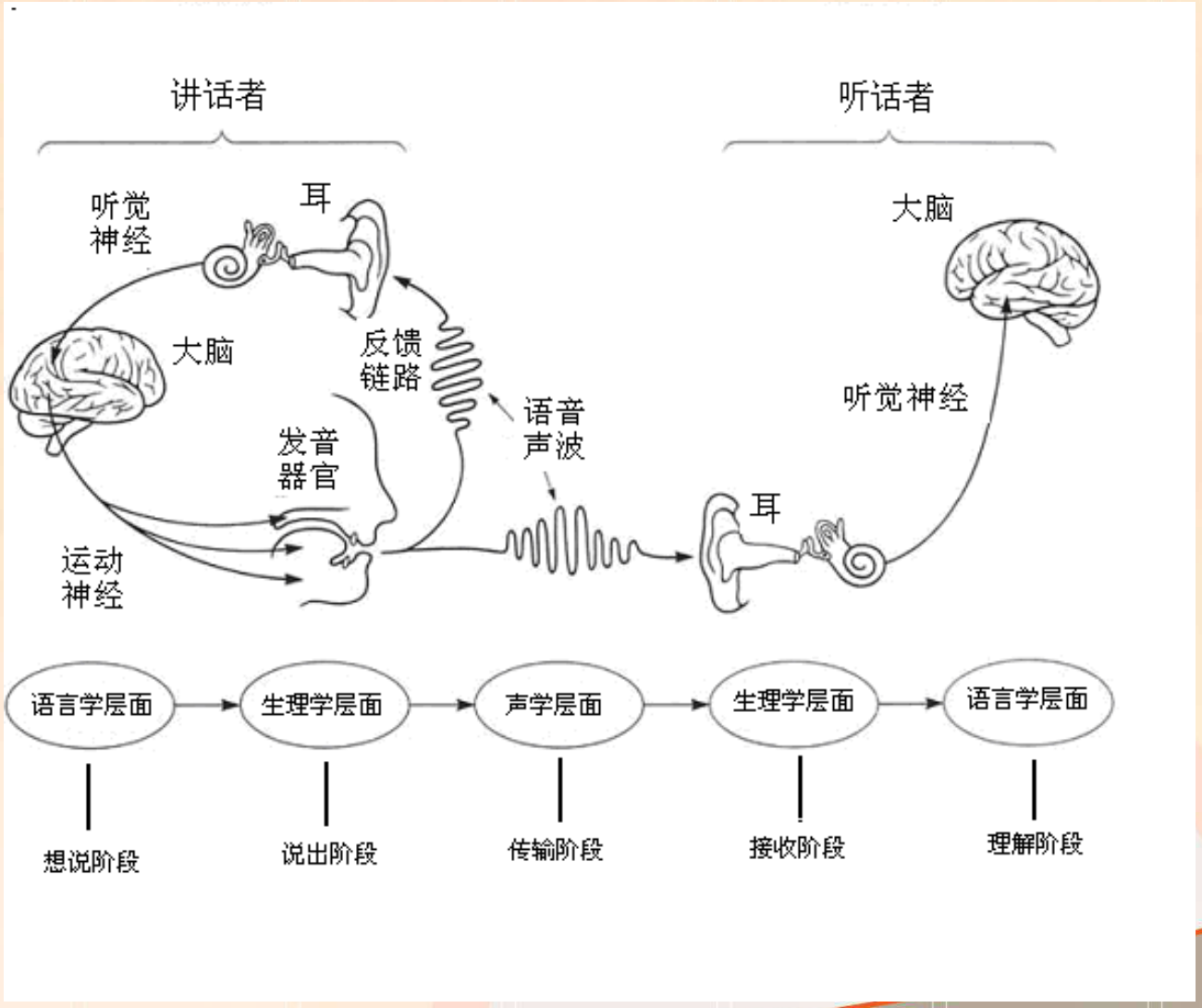
- 讲授内容
  - （9月11日）秋1：课程简介＋语音技术引言
  - （9月18日）秋2：语音分析（I）
  - （9月25日）秋3：语音分析（II）
  - （10月9日）秋5：语音识别，语音编码与合成
  - （12月30日）冬8：复习及项目成果展示（周一实验课）
- 实验内容
  - 1. PRAAT 语音分析 （9月16日）秋2
  - 2. VOICEBOX说话人识别 （9月30日）秋4

考试：2020年1月16日08:00-10:00

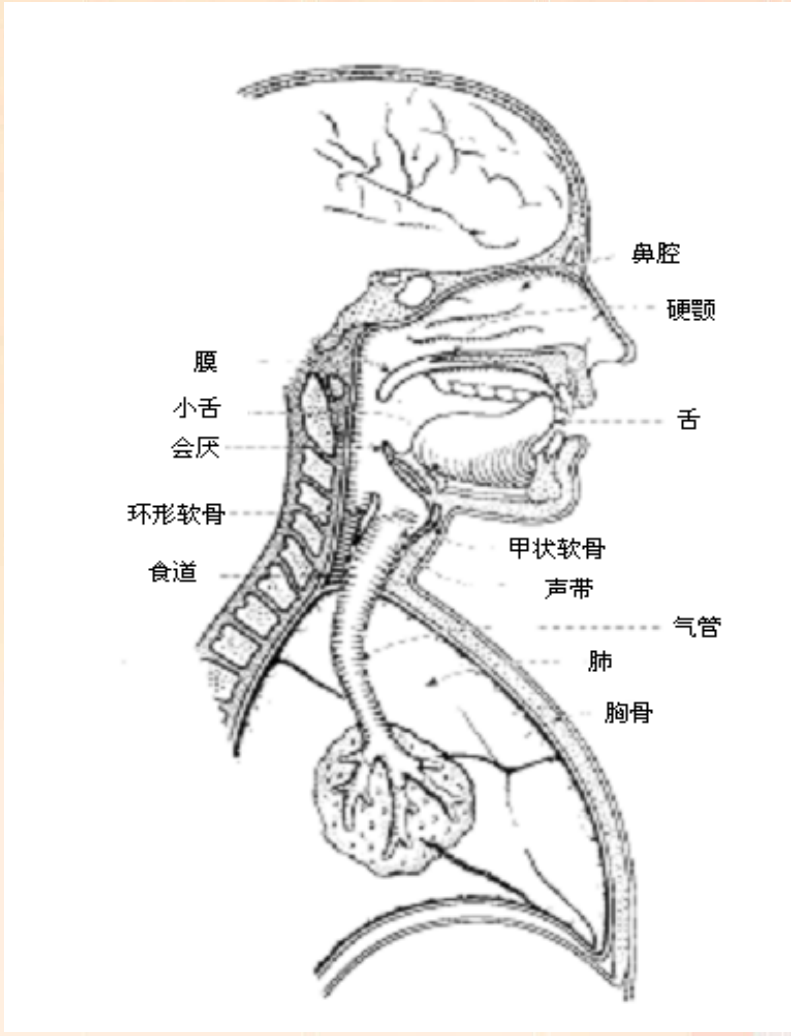
语音技术基础

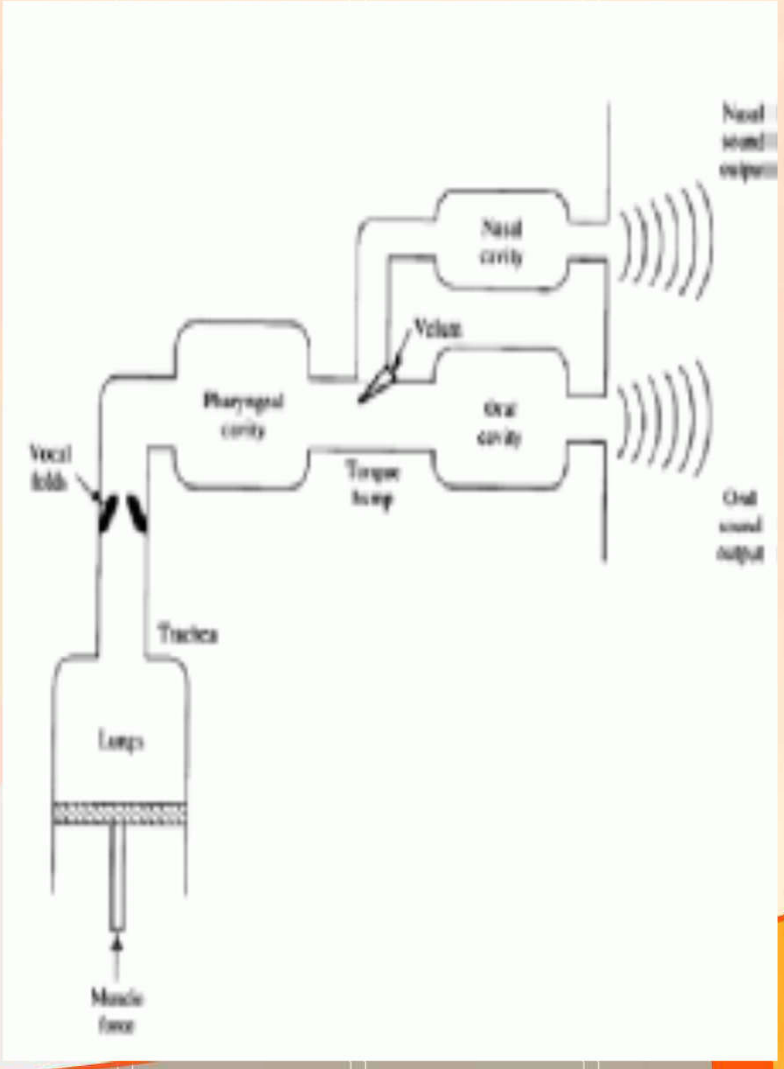
- 语言交际过程
- 语音产生过程

语音链

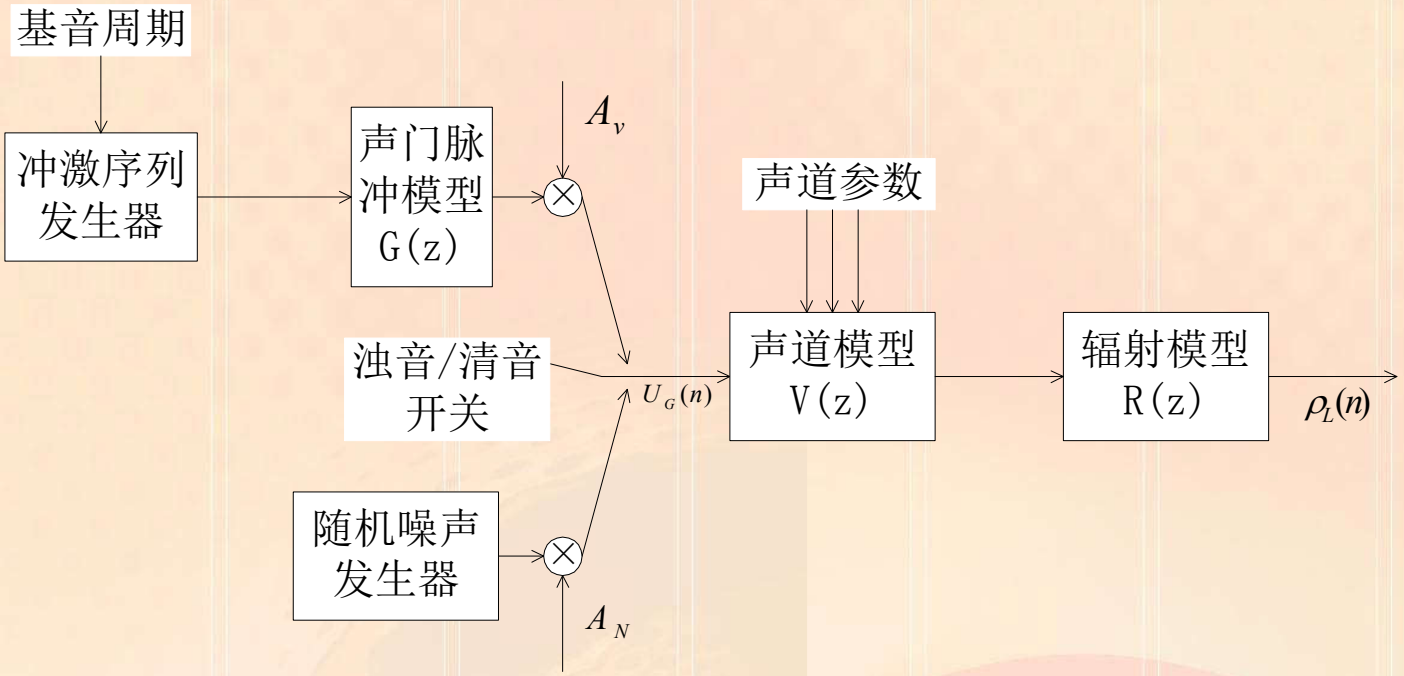

讲话者
听话者
听觉神经
耳
大脑
反馈链路
语音声波
发音器官
运动神经
大脑
听觉神经
耳
语言学层面
生理学层面
声学层面
生理学层面
语言学层面
想说阶段
说出阶段
传输阶段
接收阶段
理解阶段

语音产生模型


鼻腔
硬颚
膜
小舌
会厌
舌
环形软骨
甲状软骨
食道
声带
气管
肺
胸骨


Nasal cavity
Velum
Pharyngeal cavity
Oral cavity
Vocal folds
Tongue hump
Trachea
Lungs
Muscle force

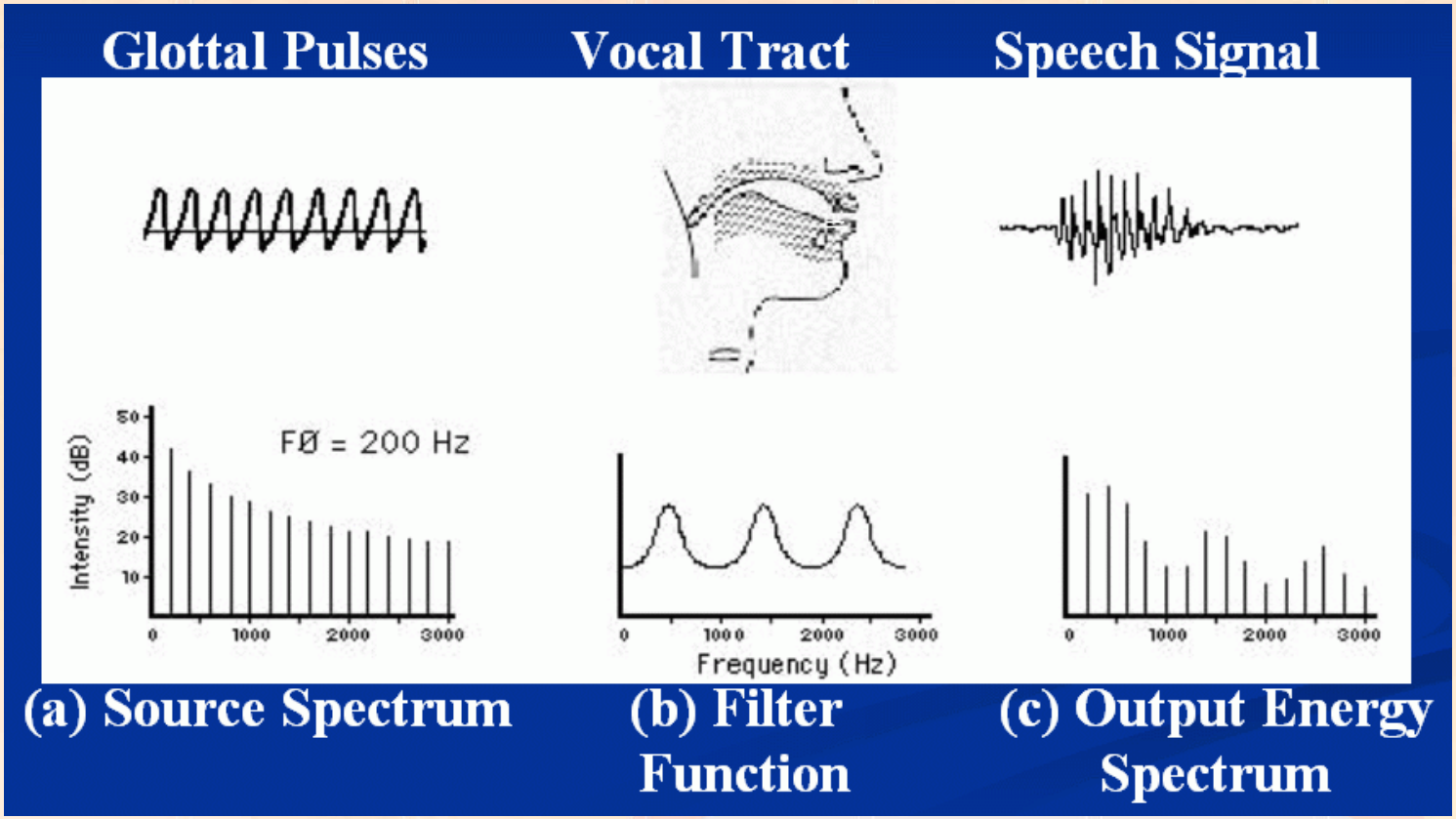
语音产生模型



语音信号产生的完整模型为

$$H(z)=U(z)V(z)R(z)$$

语音产生模型



Source-filter model and the corresponding spectrum

信号处理——表示

Transform Properties

Linearity	$ax_1[n]+bx_2[n]$	$aX_1(z)+bX_2(z)$
Shift	$x[n-n_0]$	$z^{-n_0}X(z)$
Exponential Weighting	$a^n x[n]$	$X(a^{-1}z)$
Linear Weighting	$n\,x[n]$	$-z\,dX(z)/dz$
Time Reversal	$x[-n]$ non-causal, need $x[N_0-n]$ to be causal for finite length sequence	$X(z^{-1})$
Convolution	$x[n] * h[n]$	$X(z)\,H(z)$
Multiplication of Sequences	$x[n]\,w[n]$	$\frac{1}{2\pi j}\oint_C X(v)W(z/v)v^{-1}dv$ circular convolution in the frequency domain

语音表示

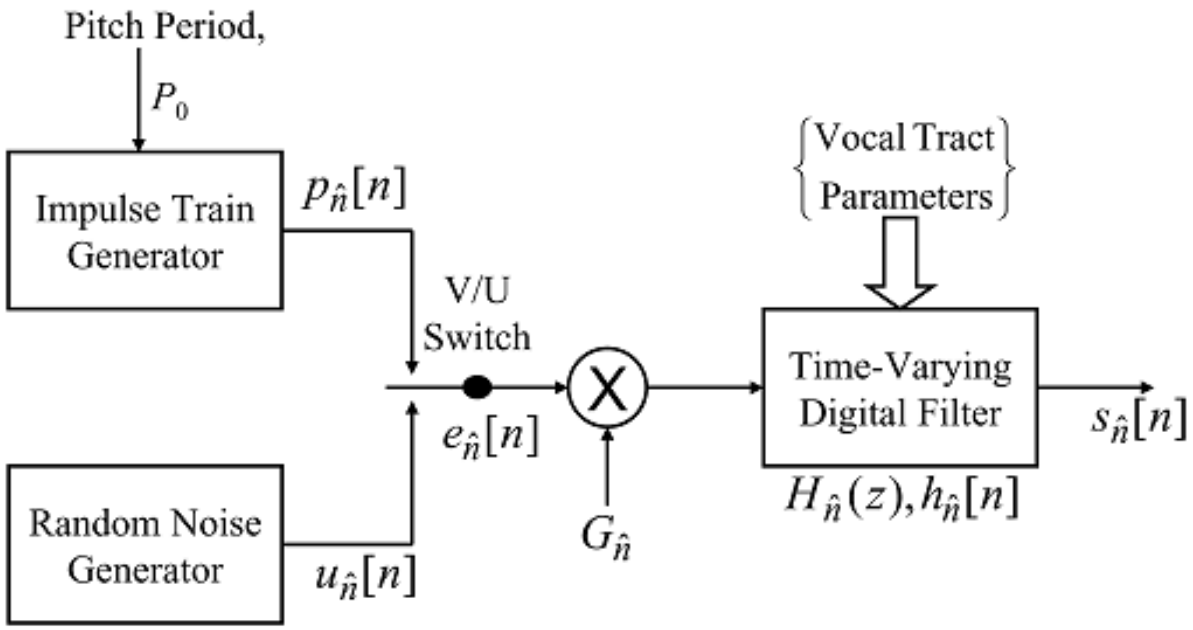


Fig. 4.1 Voiced/unvoiced/system model for a speech signal.

$$s_{\hat{n}}[n] = \sum_{m=0}^{\infty} h_{\hat{n}}[m] e_{\hat{n}}[n-m], \qquad (4.1)$$

语音表示

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}. \tag{4.2}$$

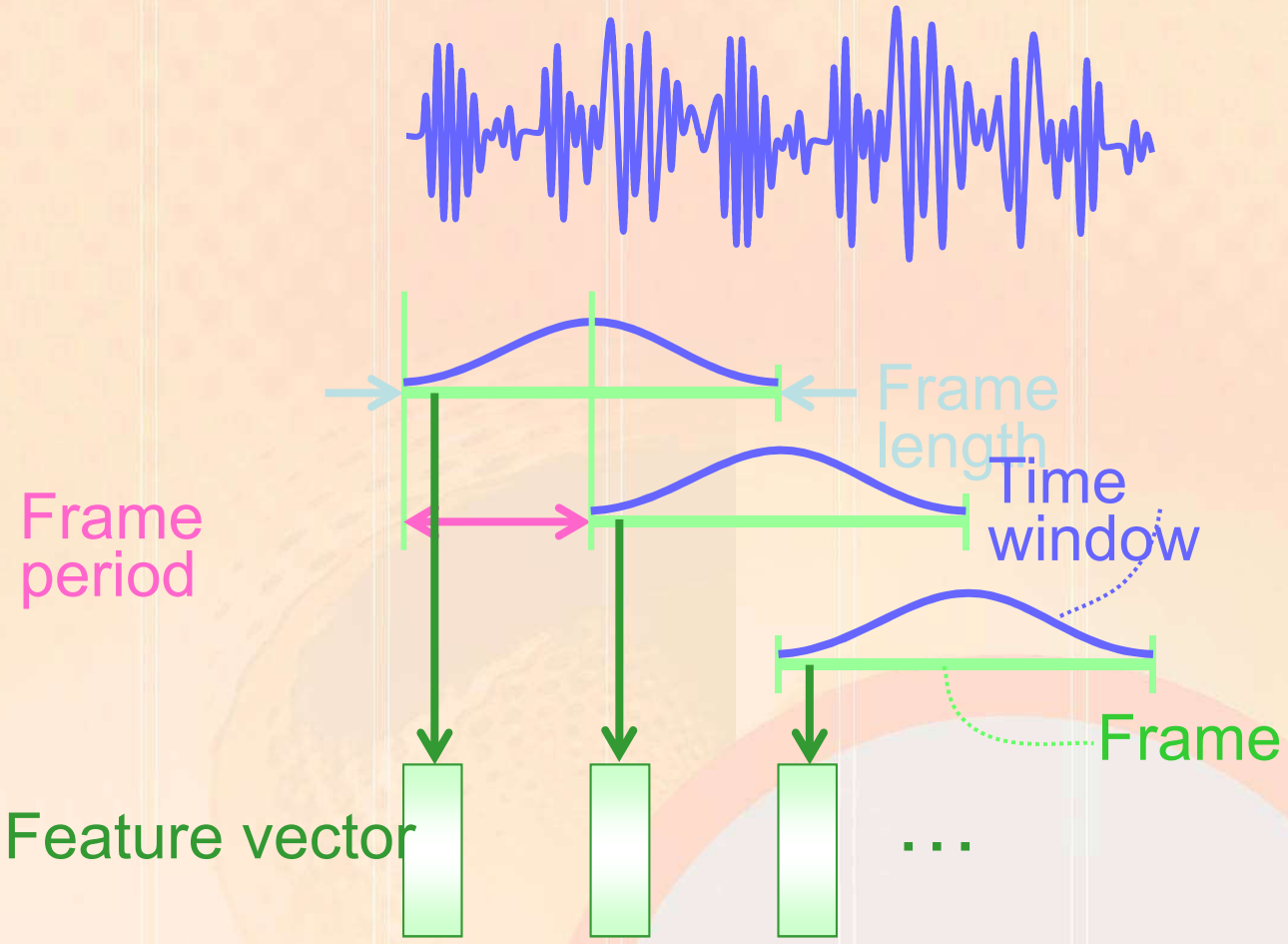
$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Ge[n], \tag{4.3}$$

$$X_{\hat{n}} = \sum_{m=-\infty}^{\infty} T\{x[m]w[\hat{n}-m]\}, \tag{4.4}$$

短时平稳假设

语音信号特性是随时间而变化的，本质上是一个非平稳过程。但不同的语音是由人的口腔肌肉运动构成声道的某种形状而产生的响应，而这种肌肉运动频率相对于语音频率来说是缓慢的，因而在一个短时间范围内，其特性基本保持不变，即相对稳定，可以视作一个准稳态过程。基于这样的考虑，对语音信号进行分段考虑，每一段称为一帧（frame）.一般假设为10-30ms的短时间隔。

语音特征提取



语音特征矢量(短时谱)提取

语音分析技术

- 语音时域分析
- 语音频域分析

Summary of Simple Time Domain Measures

$$Q_{\hat{n}} = \sum_{m=-\infty}^{\infty} T(x[m])\tilde{w}[\hat{n}-m]$$

1. Energy:

$$E_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} x^2[m]\tilde{w}[\hat{n}-m]$$

- can downsample $E_{\hat{n}}$ at rate commensurate with window bandwidth

2. Magnitude:

$$M_{\hat{n}} = \sum_{m=\hat{n}-L+1}^{\hat{n}} |x[m]|\tilde{w}[\hat{n}-m]$$

3. Zero Crossing Rate:

$$Z_{\hat{n}} = z_1 = \frac{1}{2L}\sum_{m=\hat{n}-L+1}^{\hat{n}} |\mathrm{sgn}(x[m]) - \mathrm{sgn}(x[m-1])|\tilde{w}[\hat{n}-m]$$

where $\mathrm{sgn}(x[m]) = 1 \quad x[m] \geq 0$

$= -1 \quad x[m] < 0$

语音信号时域分析

- 预处理
- 能量/音量（Energy/Volume ）
- 过零率（Zero Crossing Rate ）
- 端点检测（End-Point Detection ）
- 基频 （F0）

端点检测

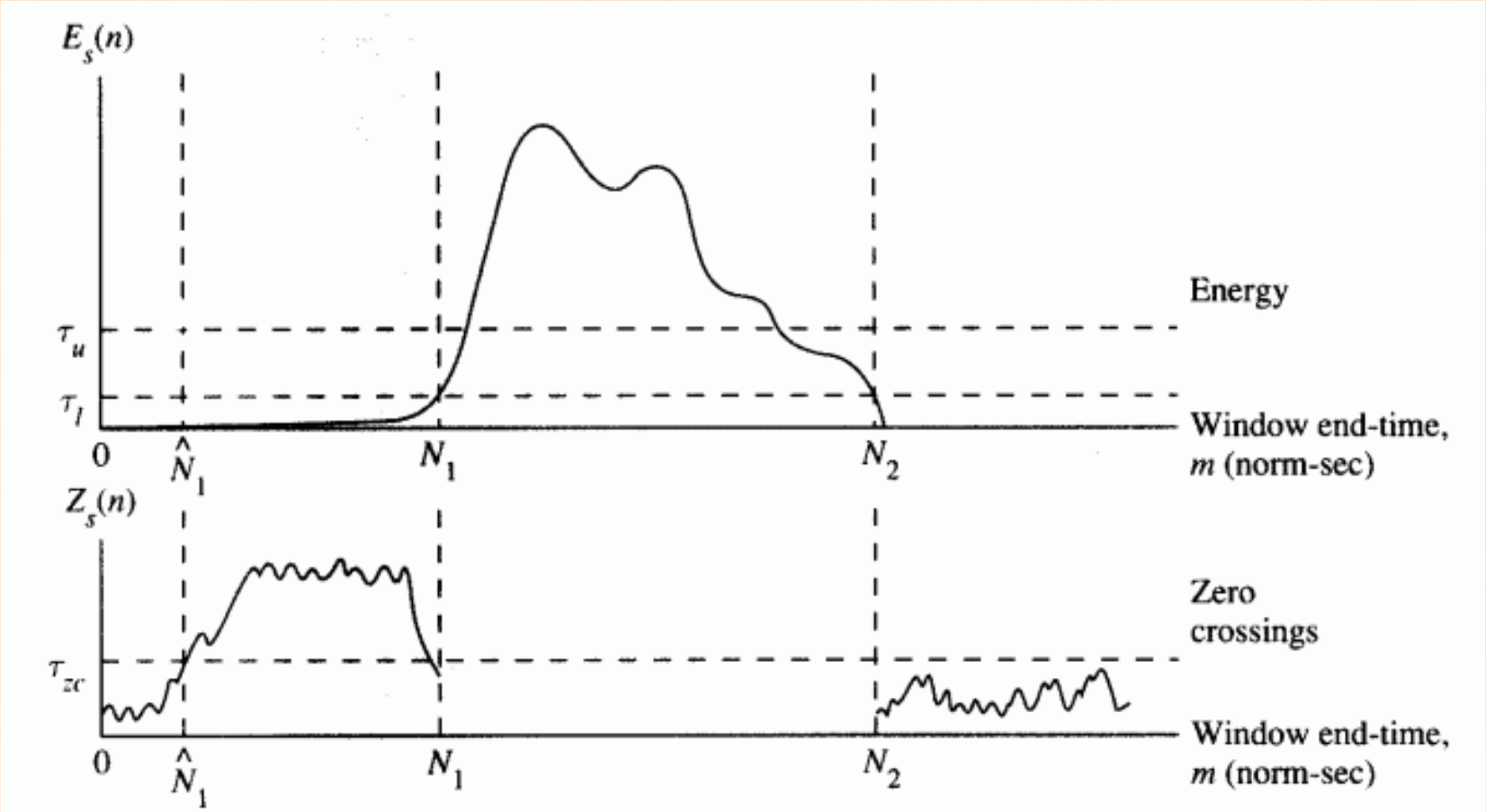
“端点检测”（End-point Detection，EPD）的目标是检测语音的开始与结束的位置，又称为 Speech Detection 或是 VAD (Voice Activity Detection)。

端点检测出错，在语音识别上会造成不良后果：

False Rejection：将Speech 误认为 Silence/Noise，造成语音识别率下降

False Acceptance：将Silence/Noise误认为 Speech，造成语音识别率下降。

端点检测


$E_s(n)$
τ_u
τ_l
0
$\hat{N}_1$
N_1
N_2
Energy
Window end-time,
m (norm-sec)
$Z_s(n)$
τ_{zc}
0
$\hat{N}_1$
N_1
N_2
Zero
crossings
Window end-time,
m (norm-sec)

端点检测

Algorithm **for endpoint detection:**

1. compute mean and σ of E_n and Z_n for first 100 msec of signal (assuming no speech in this interval).
2. determine maximum value of E_n for entire recording.
3. compute E_n thresholds based on results of steps 1 and 2—e.g., take some percentage of the peaks over the entire interval. Use threshold for zero crossings based on ZC distribution for unvoiced speech.
4. find an interval of E_n that exceeds a high threshold ITU.
5. find a putative starting point (N_1) where E_n crosses ITL from below; find a putative ending point (N_2) where E_n crosses ITL from above.
6. move backwards from N_1 by comparing Z_n to IZCT, and find the first point where Z_n exceeds IZCT; similarly move forward from N_2 by comparing Z_n to IZCT and finding last point where Z_n exceeds IZCT.

语音信号时域分析

- 预处理
- 能量/音量（Energy/Volume ）
- 过零率（Zero Crossing Rate ）
- 端点检测（End-Point Detection ）
- 基频 （F0）

auto-correlation function (ACF)

This is a time-domain method which estimates the similarity between a frame $S(i), i = 0, \cdots n-1$

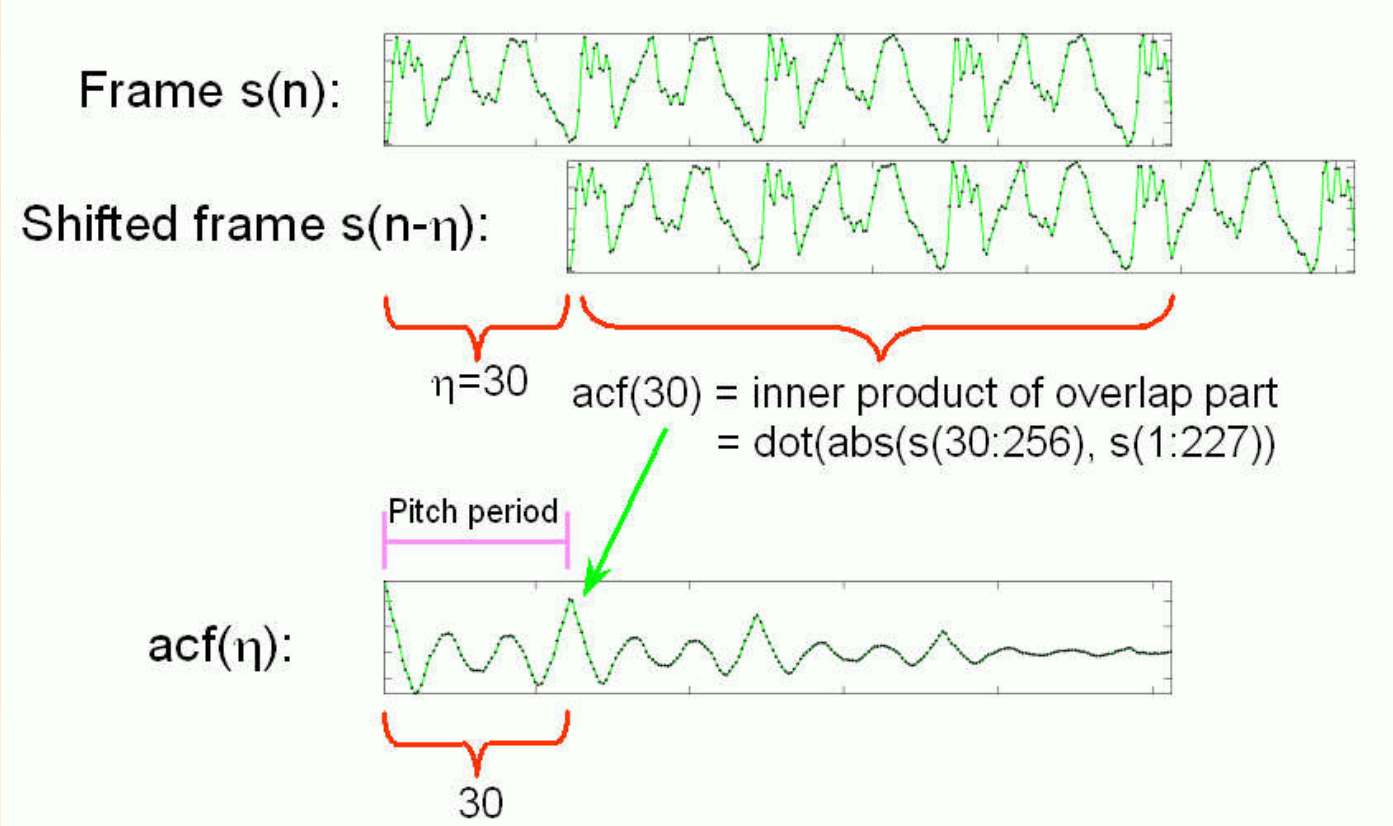
and its delayed version via the auto-correlation function:

$$acf(\tau) = \sum_{i=0}^{n-1-\tau} S(i) \bullet S(i+\tau)$$

where τ is the time lag in terms of sample points.

The value of τ that maximizes $acf(\tau)$ over a specified range is selected as the pitch period in sample points.

基频——自相关法



In other words, we shift the delayed version n times and compute the inner product of the overlapped parts to obtain n values of ACF.

参考文献

1.吴朝晖，杨莹春，说话人识别模型与方法，清华大学出版社，2009, 2

2.杨莹春，陈华，吴飞，视音频信号处理，浙江大学出版社，待出版

3. Roger Jang (張智星)

Audio Signal Processing and Recognition (音訊處理與辨識)
http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/index.asp

课后任务

- 阅读文献
  - L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing

    Ch1_Introduction

    Ch2_The Speech Signal

    Ch4_Short-Time Analysis of Speech (Pre)

语音分析技术

- 语音时域分析
- 语音频域分析

语音分析技术

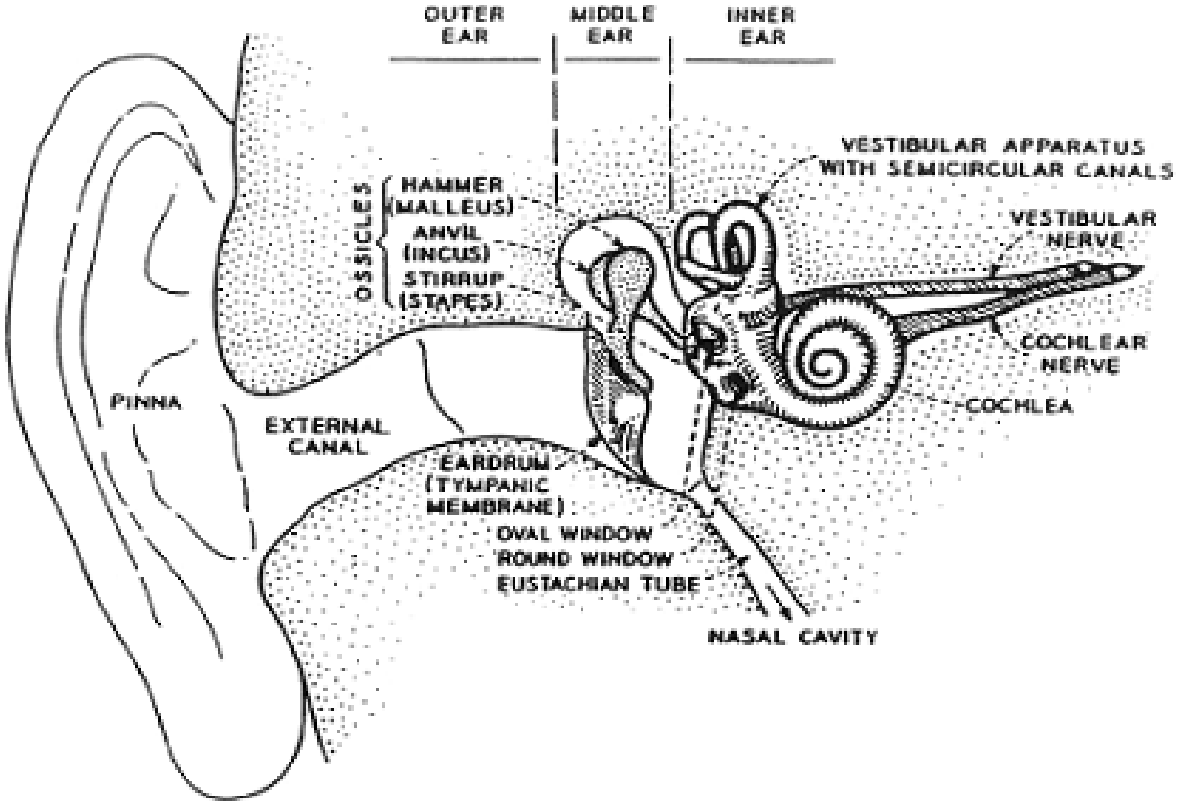
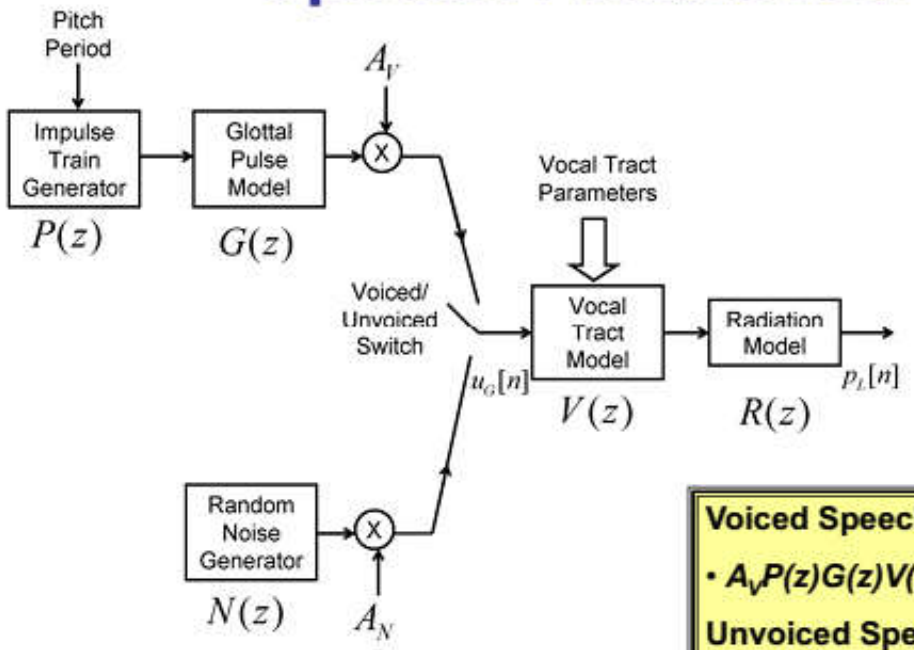


Fig. 3.1 Schematic view of the human ear (inner and middle structures enlarged). (After Flanagan [34].)

Short-Time Fourier Analysis短时傅里叶分析

General Discrete-Time Model of Speech Production



Voiced Speech:

- $A_V P(z)G(z)V(z)R(z)$

Unvoiced Speech:

- $A_N N(z)V(z)R(z)$

Short-Time Fourier Analysis短时傅里叶分析

Short-Time Fourier Analysis

- represent signal by **sum of sinusoids** or complex exponentials as it leads to convenient solutions to problems (formant estimation, pitch period estimation, analysis-by-synthesis methods), and insight into the signal itself
- such **Fourier representations** provide
  - convenient means to determine response to a sum of sinusoids for linear systems
  - clear evidence of signal properties that are obscured in the original signal

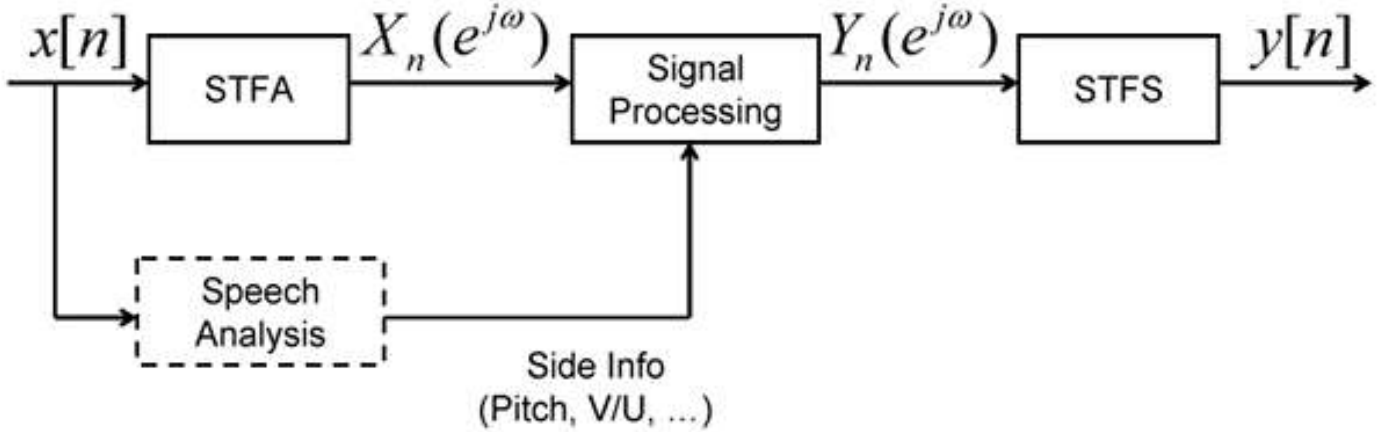
Short-Time Fourier Analysis短时傅里叶分析

Why STFT for Speech Signals

- steady state sounds, like vowels, are produced by **periodic excitation of a linear system** => speech spectrum is the product of the excitation spectrum and the vocal tract frequency response
- speech is a **time-varying signal** => need more sophisticated analysis to reflect time varying properties
  - changes occur at syllabic rates (~10 times/sec)
  - over fixed time intervals of 10-30 msec, properties of most speech signals are relatively constant (when is this not the case)

Short-Time Fourier Analysis短时傅里叶分析

Frequency Domain Processing



- **Coding**:
  - transform, subband, homomorphic, channel vocoders
- **Restoration/Enhancement/Modification**:
  - noise and reverberation removal, helium restoration, time-scale modifications (speed-up and slow-down of speech)

Short-Time Fourier Analysis短时傅里叶分析

Frequency and the *DTFT*

- sinusoids

$$x(n) = \cos(\omega_0 n) = (e^{j\omega_0 n} + e^{-j\omega_0 n})/2$$

where ω_0 is the *frequency* (in radians) of the sinusoid

- the Discrete-Time Fourier Transform (*DTFT*)

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n} = DTFT\{x(n)\}$$

$$x(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega = DTFT^{-1}\{X(e^{j\omega})\}$$

where ω is the *frequency variable* of $X(e^{j\omega})$

Short-Time Fourier Analysis短时傅里叶分析

DTFT and DFT of Speech

- The DTFT and the DFT for the infinite duration signal could be calculated (the DTFT) and approximated (the DFT) by the following:

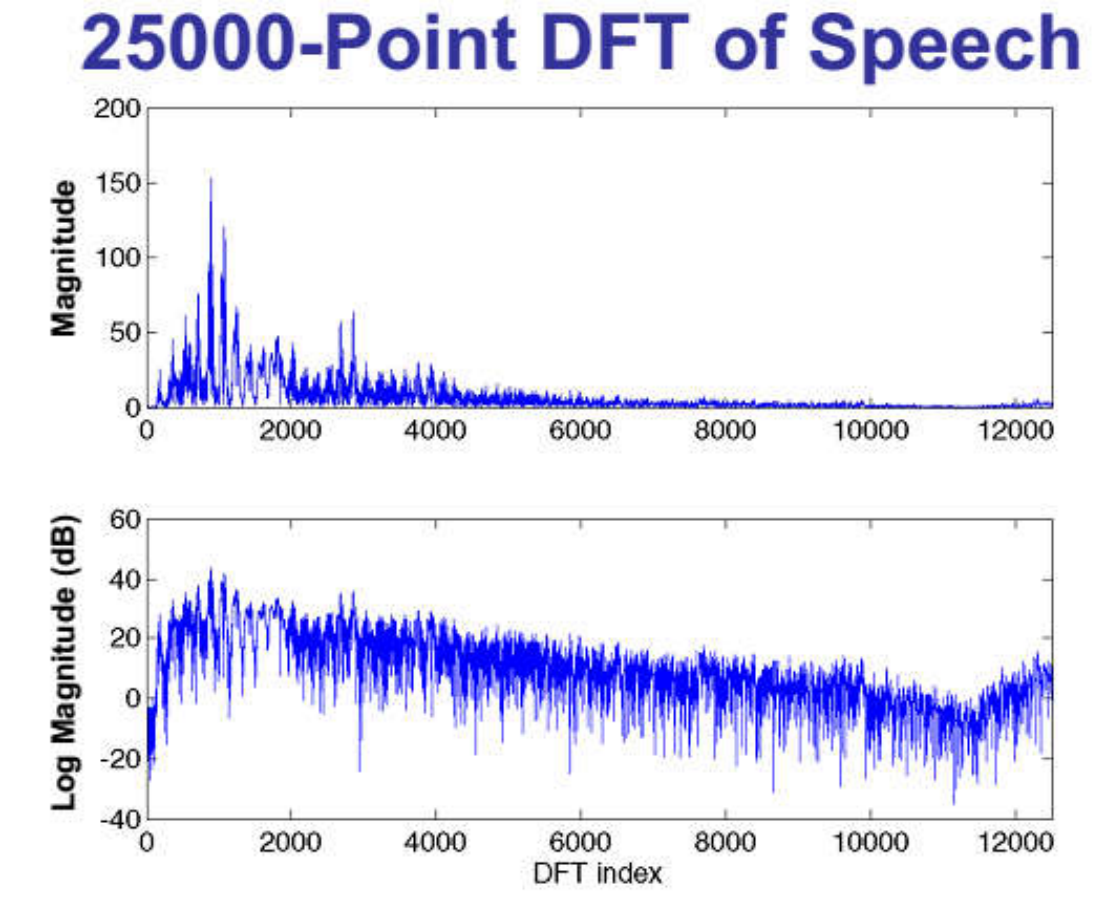
$$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\omega m} \quad (DTFT)$$

$$X(k) = \sum_{m=0}^{L-1} x(m)w(m)e^{-j(2\pi/L)km}, \quad k = 0,1,\ldots,L-1$$

$$= X(e^{j\omega})\Big|_{\omega=(2\pi k/L)} \quad (DFT)$$

- using a value of L=25000 we get the following plot

Short-Time Fourier Analysis短时傅里叶分析



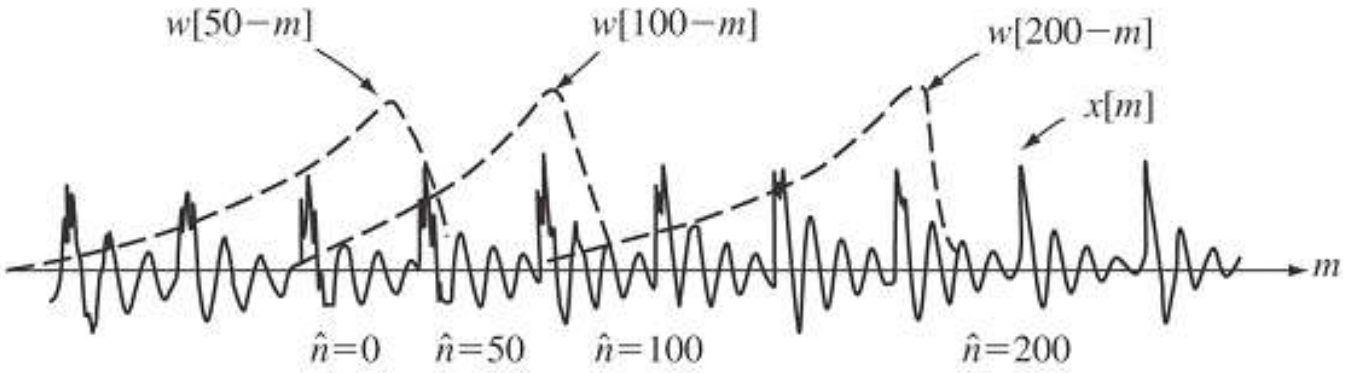
Short-Time Fourier Analysis短时傅里叶分析

Definition of STFT

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(m)w(\hat{n}-m)e^{-j\hat{\omega}m}$$

both $\hat{n}$ and $\hat{\omega}$ are variables

- $w(\hat{n}-m)$ is a real window which determines the portion of $x(\hat{n})$ that is used in the computation of $X_{\hat{n}}(e^{j\hat{\omega}})$



Short-Time Fourier Analysis短时傅里叶分析

Short-Time Fourier Transform

- alternative form of STFT (based on change of variables) is

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} w(m)x(\hat{n}-m)e^{-j\hat{\omega}(\hat{n}-m)}$$

$$= e^{-j\hat{\omega}\hat{n}} \sum_{m=-\infty}^{\infty} x(\hat{n}-m)w(m)e^{j\hat{\omega}m}$$

- if we define

$$\tilde{X}_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x(\hat{n}-m)w(m)e^{j\hat{\omega}m}$$

- then $X_{\hat{n}}(e^{j\hat{\omega}})$ can be expressed as (using $m' = -m$)

$$X_{\hat{n}}(e^{j\hat{\omega}}) = e^{-j\hat{\omega}\hat{n}}\tilde{X}_n(e^{j\hat{\omega}}) = e^{-j\hat{\omega}\hat{n}}\,DTFT\left[x(\hat{n}+m)w(-m)\right]$$

Short-Time Fourier Analysis短时傅里叶分析

Frequencies for STFT

- the STFT is periodic in ω with period 2π, i.e.,
  $X_{\hat{n}}(e^{j\hat{\omega}}) = X_{\hat{n}}(e^{j(\hat{\omega}+2\pi k)}), \forall k$
- can use any of several frequency variables to express STFT, including
  --$\hat{\omega} = \hat{\Omega}T$ (where T is the sampling period for $x(m)$) to represent analog radian frequency, giving $X_{\hat{n}}(e^{j\hat{\Omega}T})$
  --$\hat{\omega} = 2\pi\hat{f}$ or $\hat{\omega} = 2\pi\hat{F}T$ to represent normalized frequency ($0 \le \hat{f} \le 1$) or analog frequency ($0 \le \hat{F} \le F_s = 1/T$), giving $X_{\hat{n}}(e^{j2\pi\hat{f}})$ or $X_{\hat{n}}(e^{j2\pi\hat{F}T})$

Short-Time Fourier Analysis短时傅里叶分析

Signal Recovery from STFT

- since for a given value of $\hat{n}$, $X_{\hat{n}}(e^{j\hat{\omega}})$ has the same properties as a normal Fourier transform, we can recover the input sequence exactly
- since $X_{\hat{n}}(e^{j\hat{\omega}})$ is the normal Fourier transform of the windowed sequence $w(\hat{n}-m)x(m)$, then

$$w(\hat{n}-m)x(m)=\frac{1}{2\pi}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\hat{\omega}m}d\hat{\omega}$$

- assuming the window satisfies the property that $w(0)\neq 0$ (a trivial requirement), then by evaluating the inverse Fourier transform when $m=\hat{n}$, we obtain

$$x(\hat{n})=\frac{1}{2\pi w(0)}\int_{-\pi}^{\pi}X_{\hat{n}}(e^{j\hat{\omega}})e^{j\omega\hat{n}}d\hat{\omega}$$

Short-Time Fourier Analysis短时傅里叶分析

$$S(t_r, f_k) = 20\log_{10}|\tilde{X}_{rR}[k]| = 20\log_{10}|X_{rR}[k]|. \qquad (4.21)$$

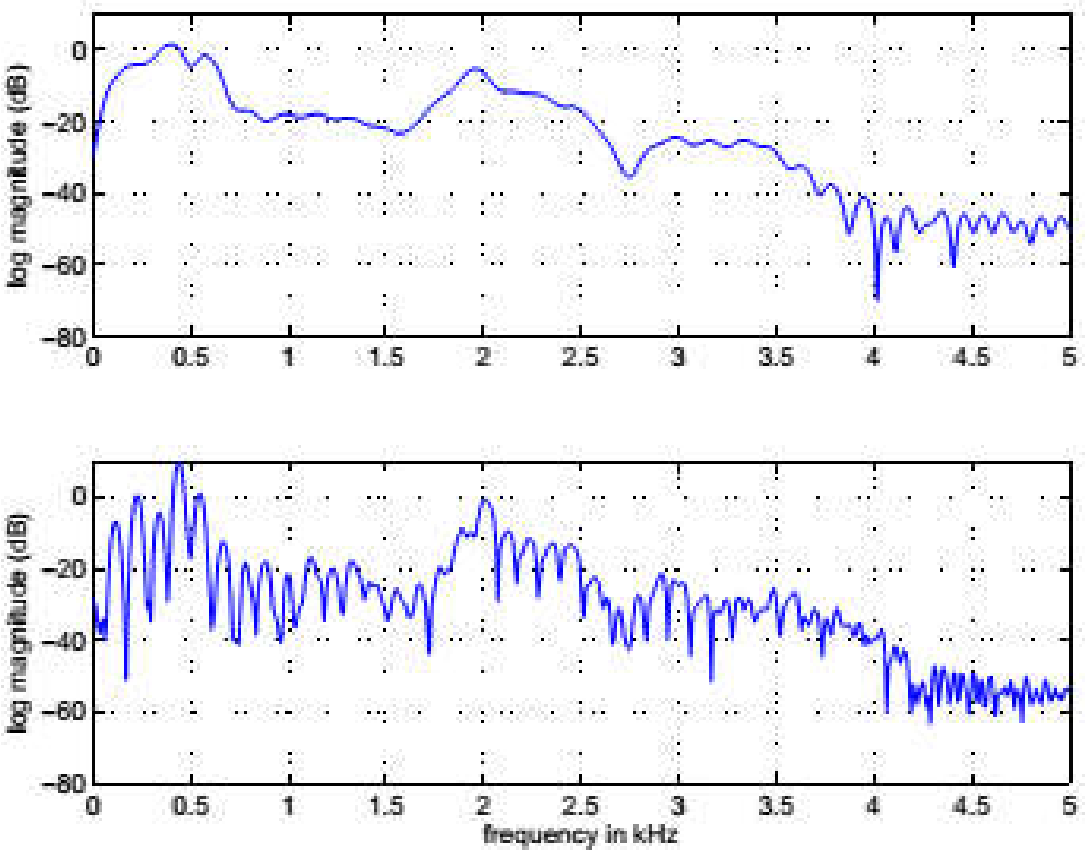


Fig. 4.7 Short-time spectrum at time 430 ms (dark vertical line in Figure 4.6) with Hamming window of length $M = 101$ in upper plot and $M = 401$ in lower plot.

Speech spectrogram

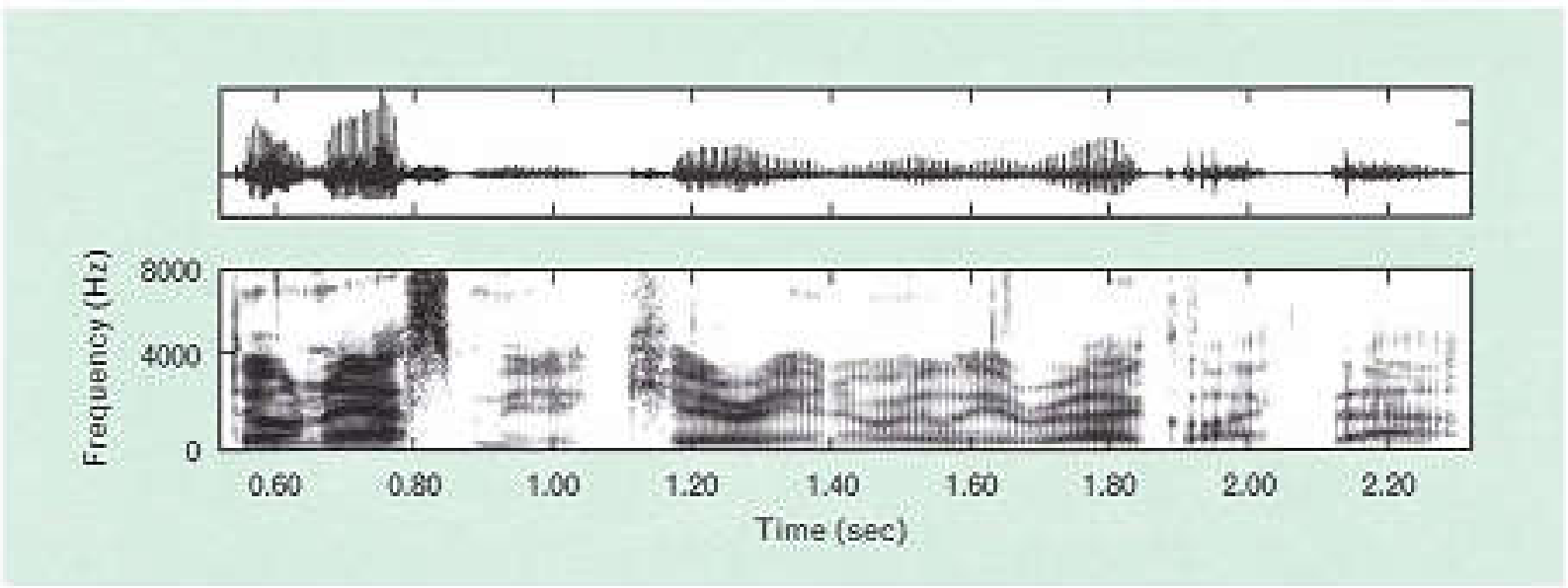
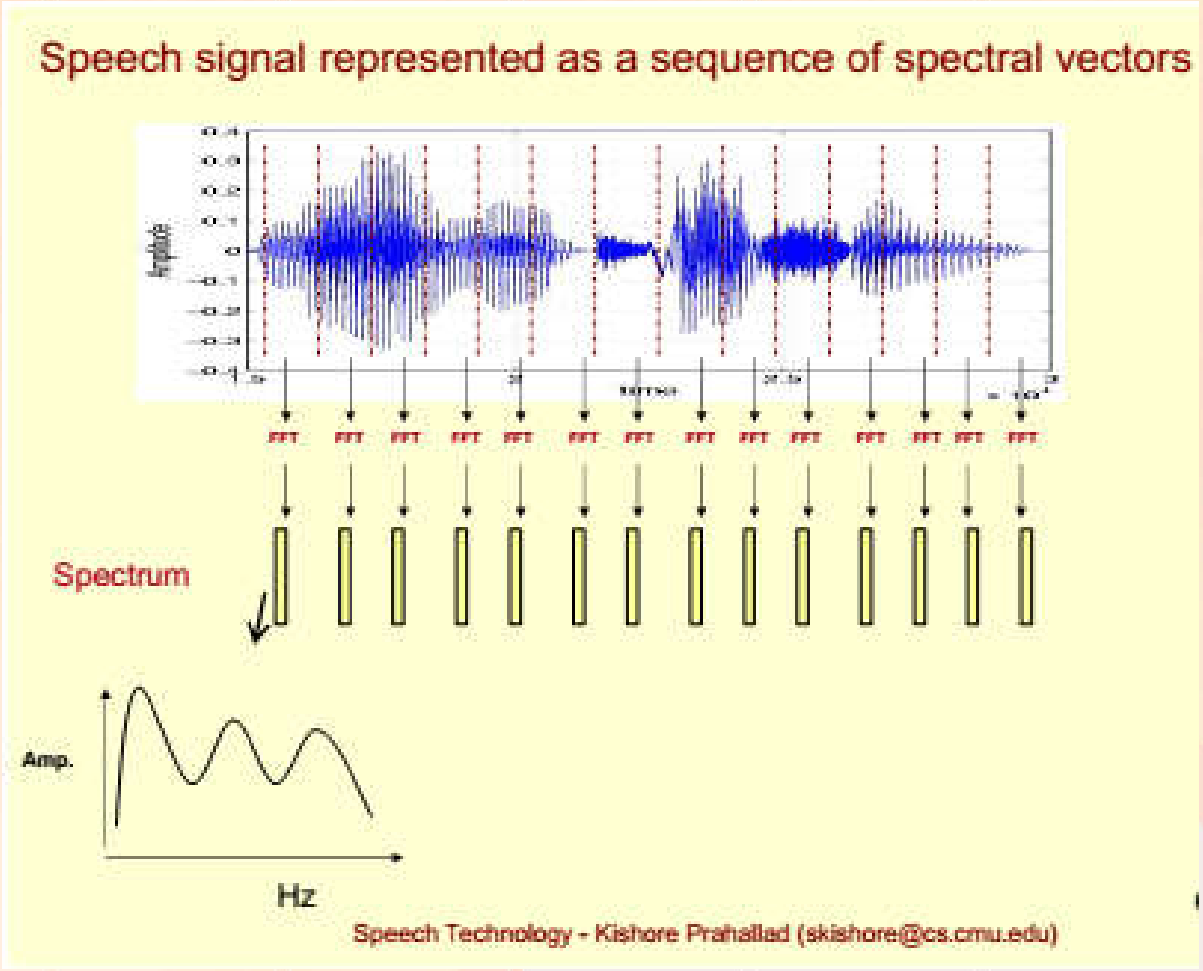
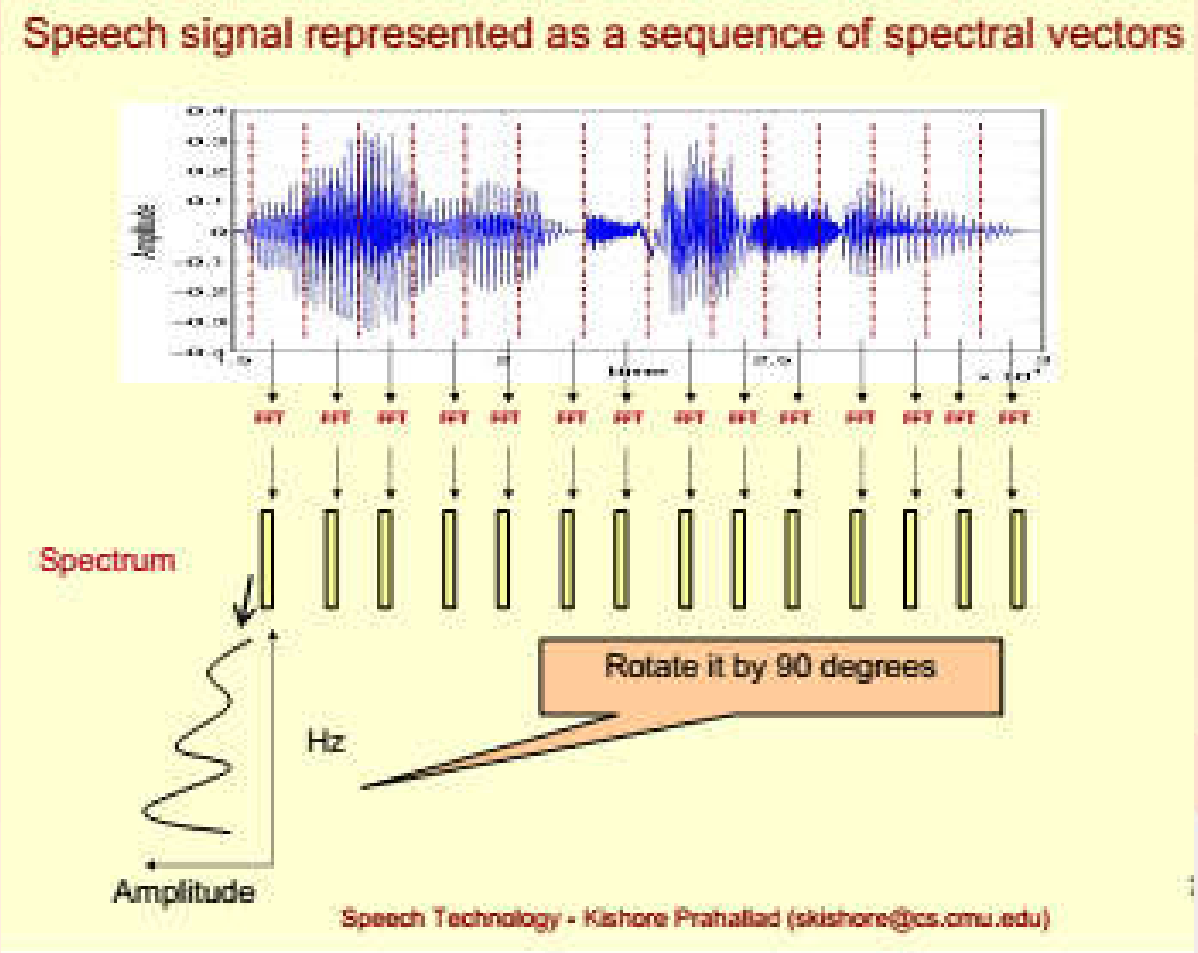


FIGURE 3. Digitally sampled speech waveform of a spoken sentence (above) and corresponding spectrogram (below) showing the dynamic nature of the formants as the vocal tract continuously changes shape. The sentence spoken was "Don't ask me to carry an oily rag like that."

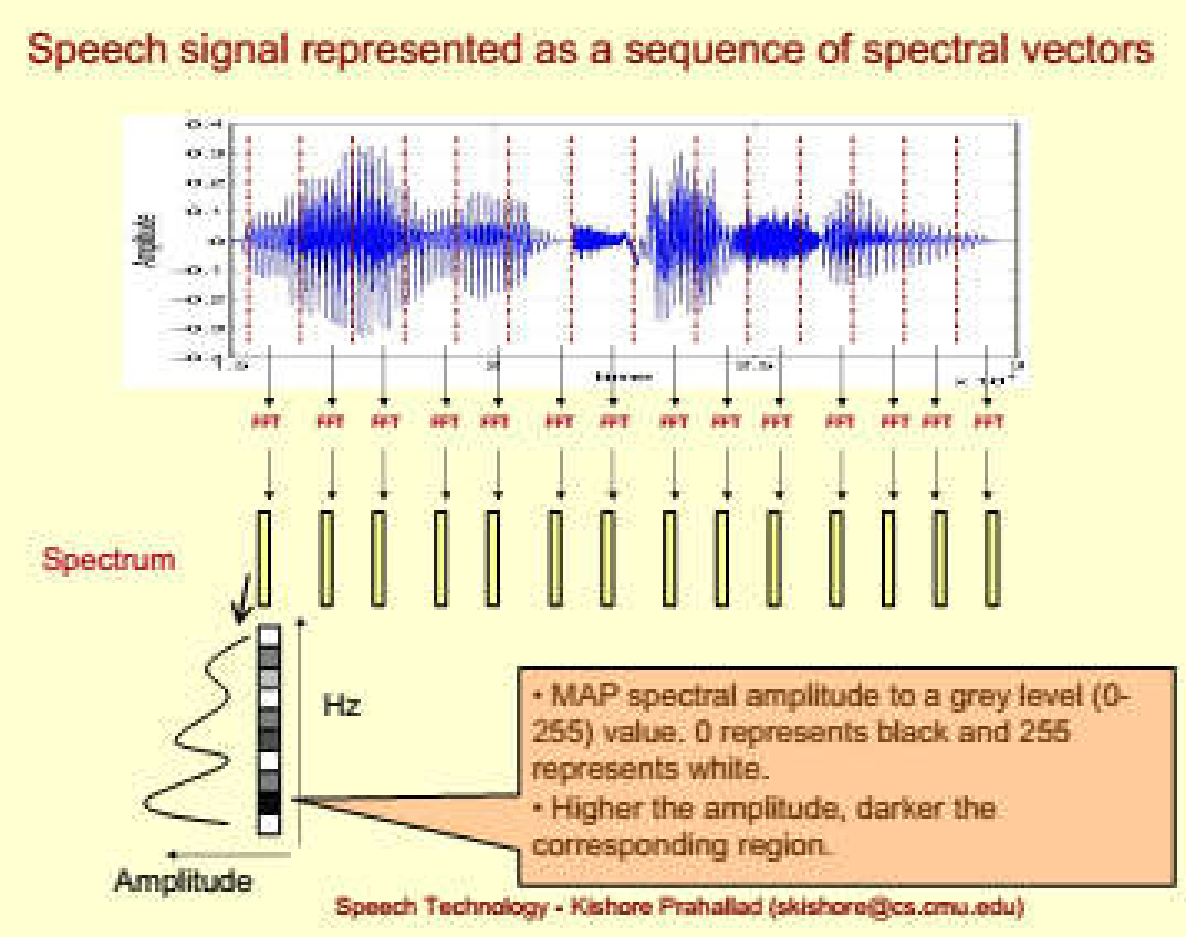
Speech spectrogram



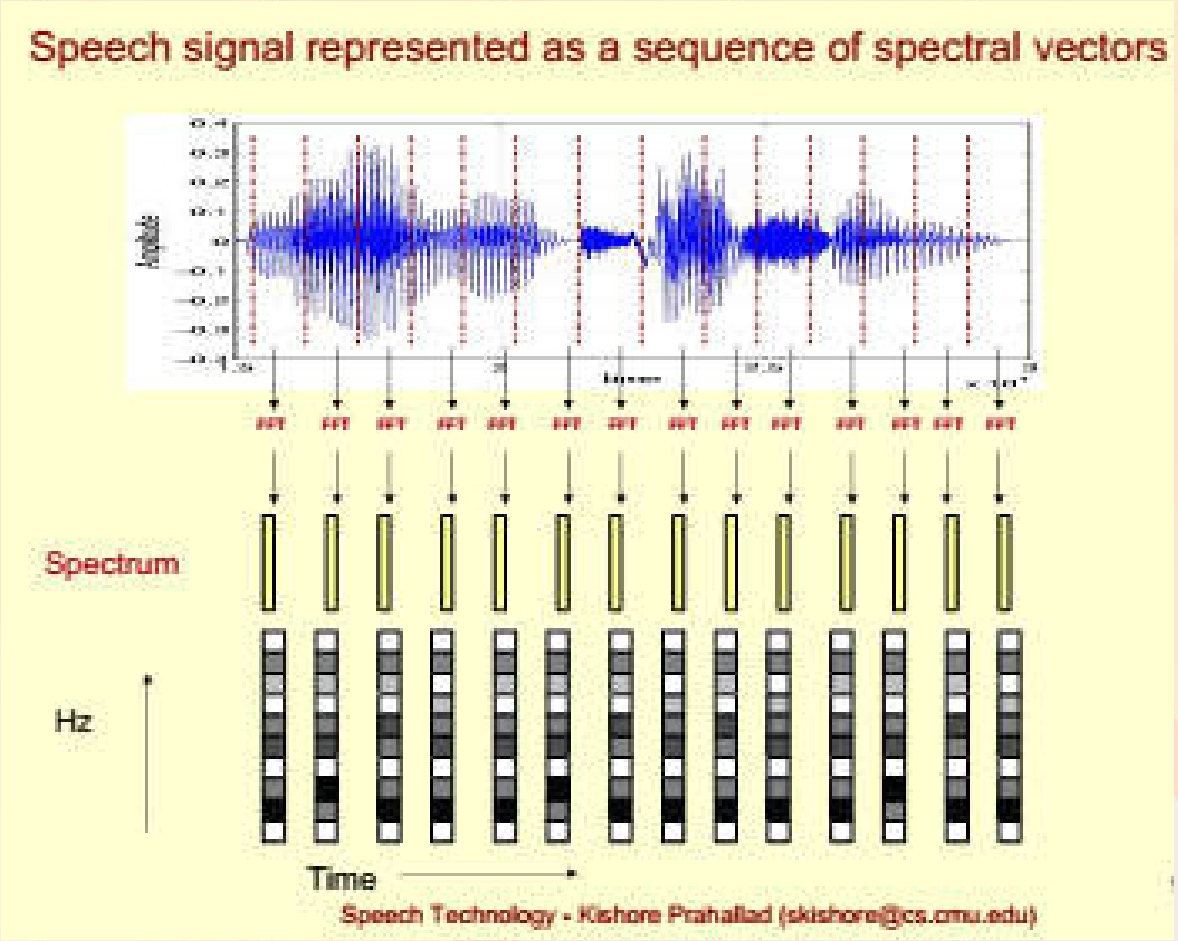
Speech spectrogram



Speech spectrogram

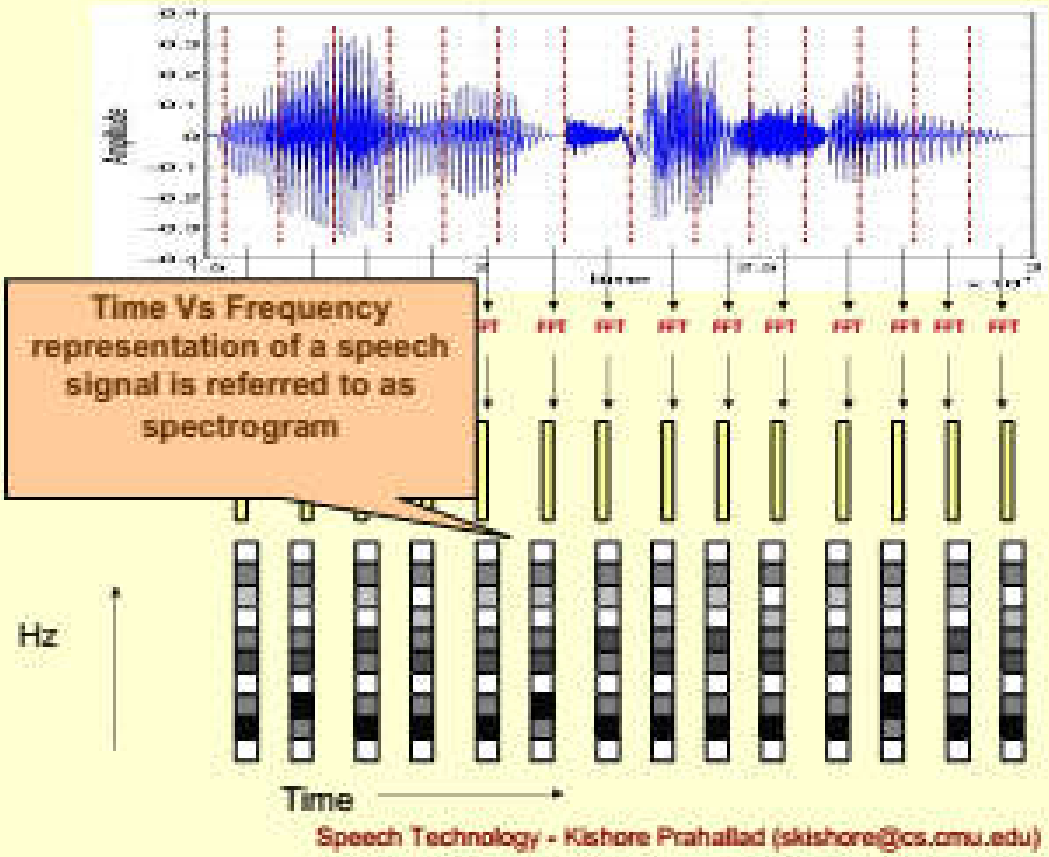


Problem Statement



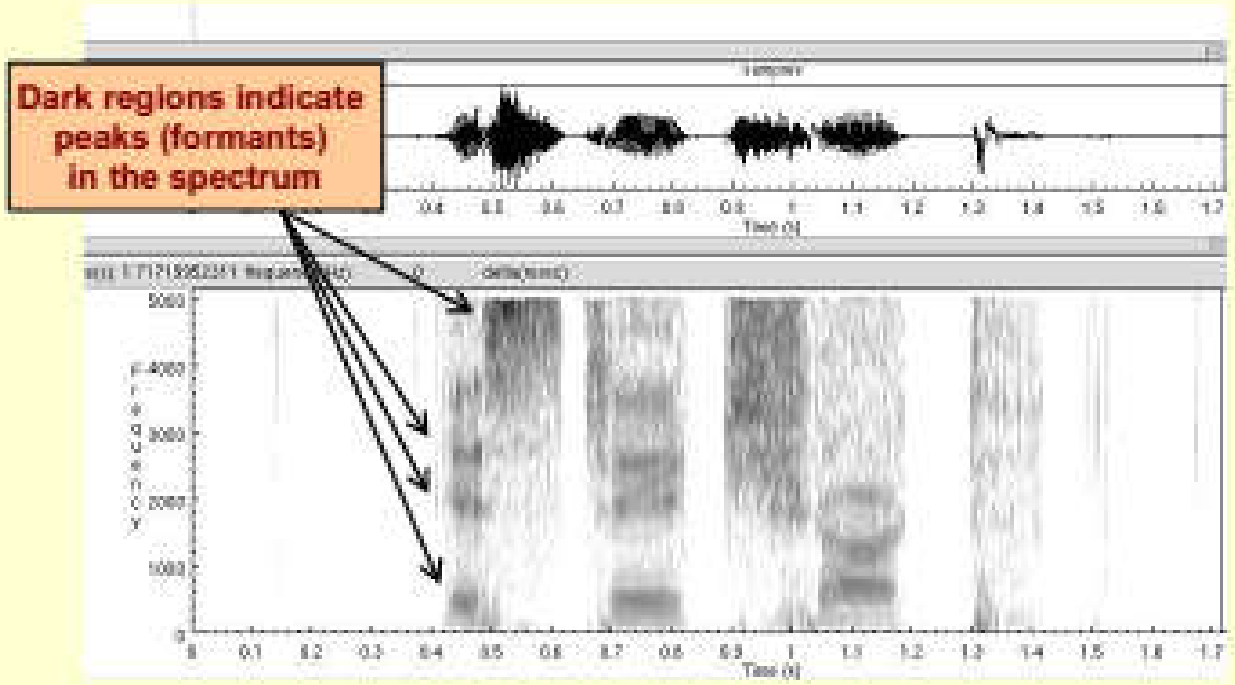
Speech spectrogram

Speech signal represented as a sequence of spectral vectors



Problem Statement

Some Real Spectrograms

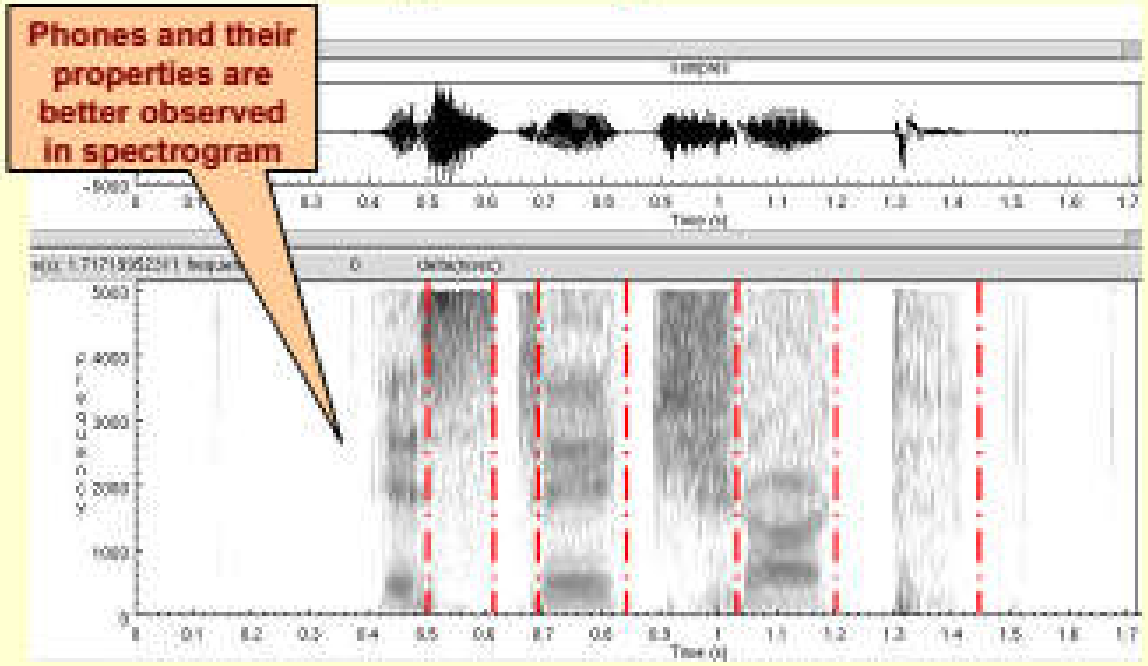


11

Speech Technology - Kishore Prahallad (skishore@cs.cmu.edu)

Speech spectrogram

Why we are bothered about spectrograms

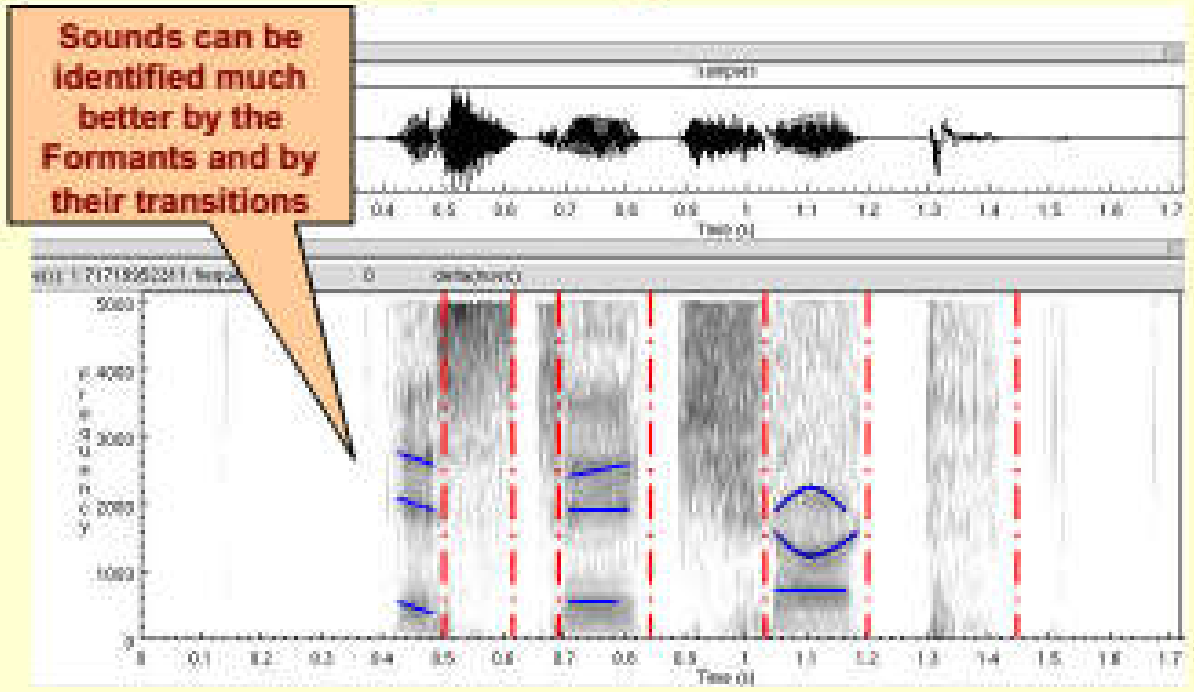


12

Speech Technology - Kishore Prahallad (skishore@cs.cmu.edu)

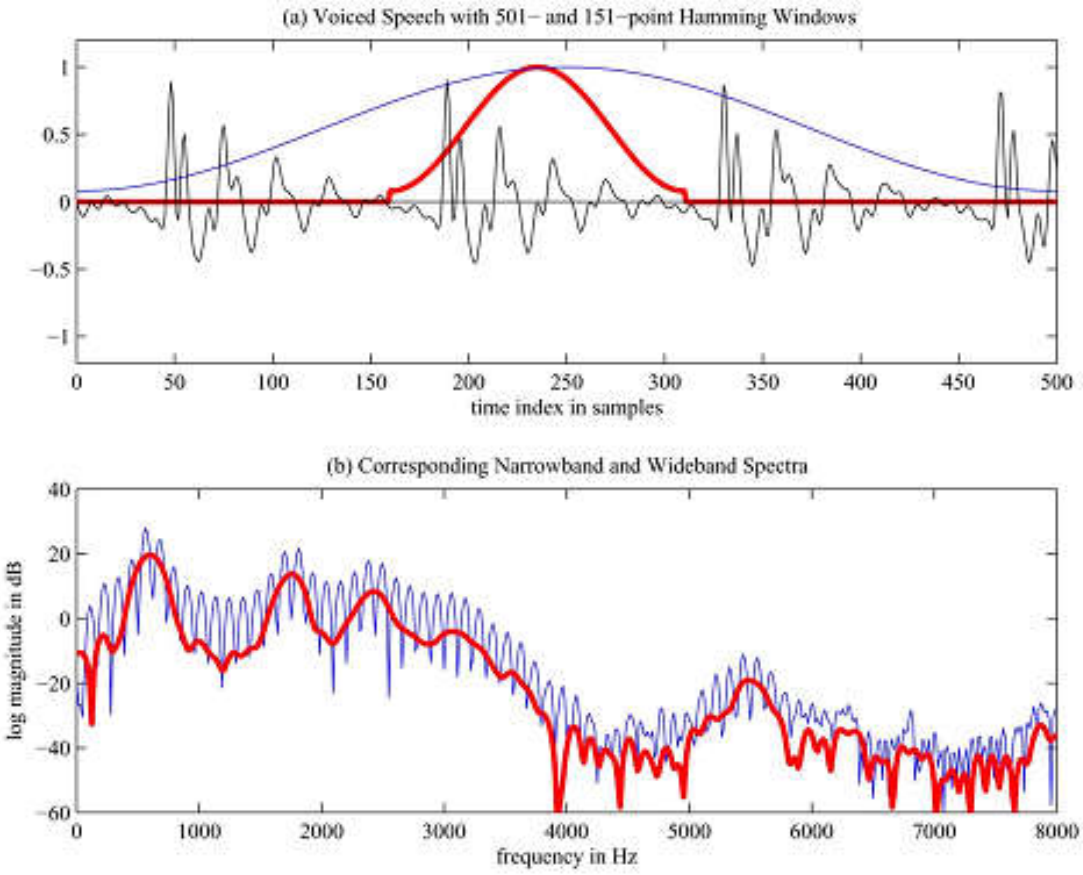
Speech spectrogram

Why we are bothered about spectrograms



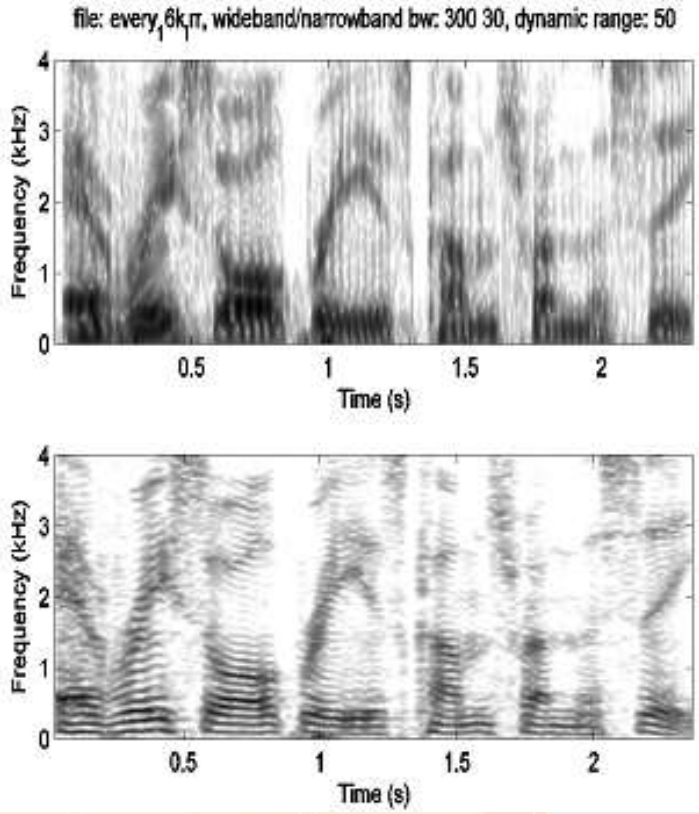
Short-Time Fourier Analysis短时傅里叶分析

Effect of Window Length-HW



speech spctrogram 语图

Digital Speech Spectrograms



- **wideband spectrogram**
  - follows broad spectral peaks (formants) over time
  - resolves most individual pitch periods as vertical striations since the IR of the analyzing filter is comparable in duration to a pitch period
  - what happens for low pitch males—high pitch females
  - for unvoiced speech there are no vertical pitch striations
- **narrowband spectrogram**
  - individual harmonics are resolved in voiced regions
  - formant frequencies are still in evidence
  - usually can see fundamental frequency
  - unvoiced regions show no strong structure

64

Cepstrum analysis

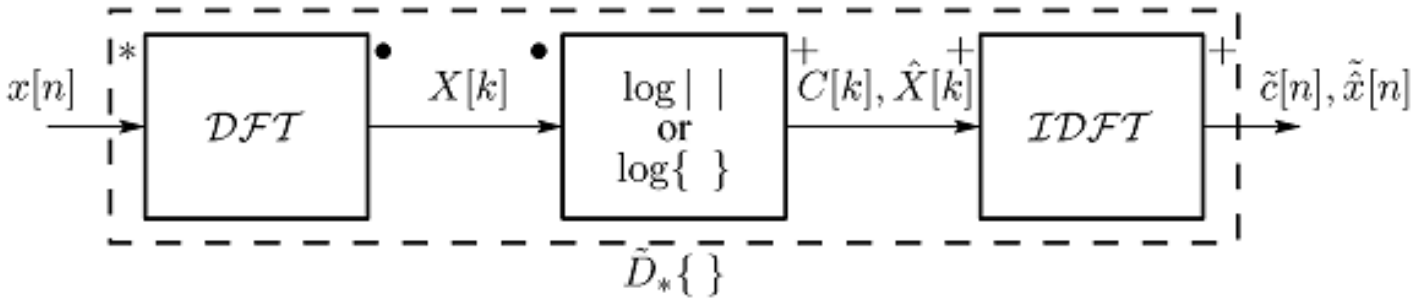
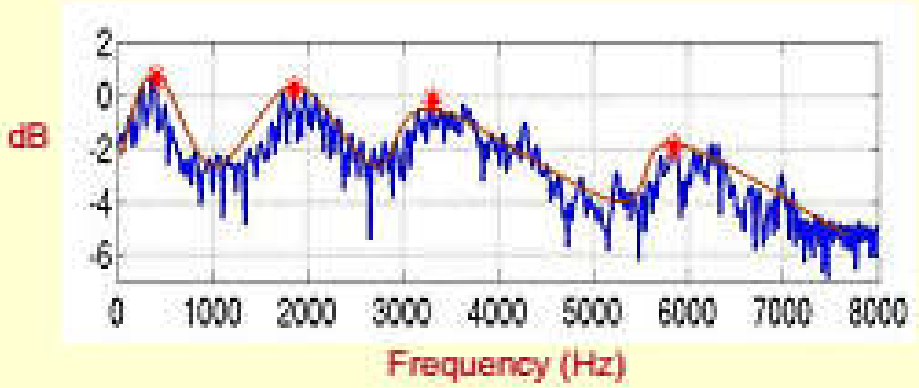


Fig. 5.3 Computing the cepstrum or complex cepstrum using the DFT.

MFCC

- We captured spectral envelope (curve connecting all formants)
- BUT: Perceptual experiments say human ear concentrates on certain regions rather than using whole of the spectral envelope….



Speech Technology - Kishore Prahallad (skishore@cs.cmu.edu)

MFCC

Mel-Frequency Analysis

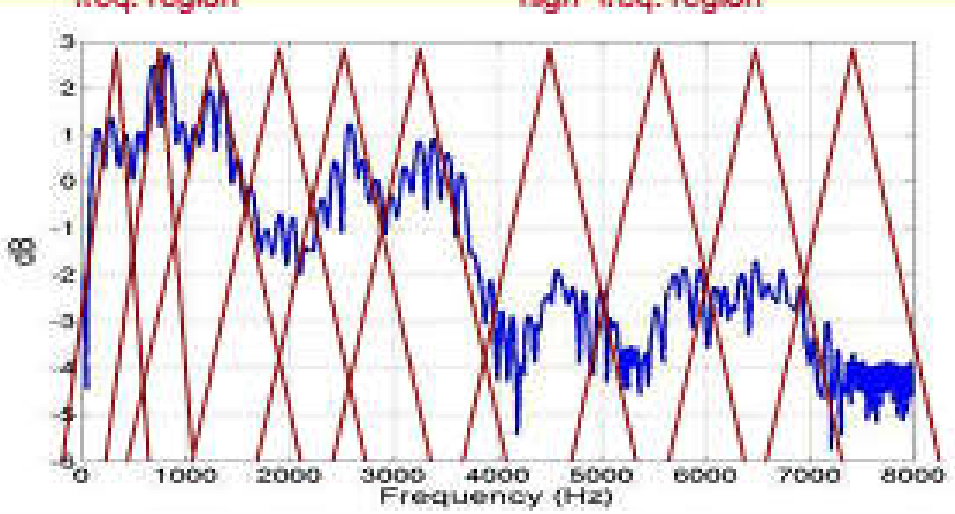
- Mel-Frequency analysis of speech is based on human perception experiments
- It is observed that human ear acts as filter
  - It concentrates on only certain frequency components
- These filters are non-uniformly spaced on the frequency axis
  - More filters in the low frequency regions
  - Less no. of filters in high frequency regions

MFCC

Mel-Frequency Filters

More no. of filters in low freq. region

Lesser no. of filters in high freq. region



MFCC

The basic idea is to compute a frequency analysis based upon a filter bank with approximately critical band spacing of the filters and bandwidths. For 4 kHz bandwidth, approximately 20 filters are used.

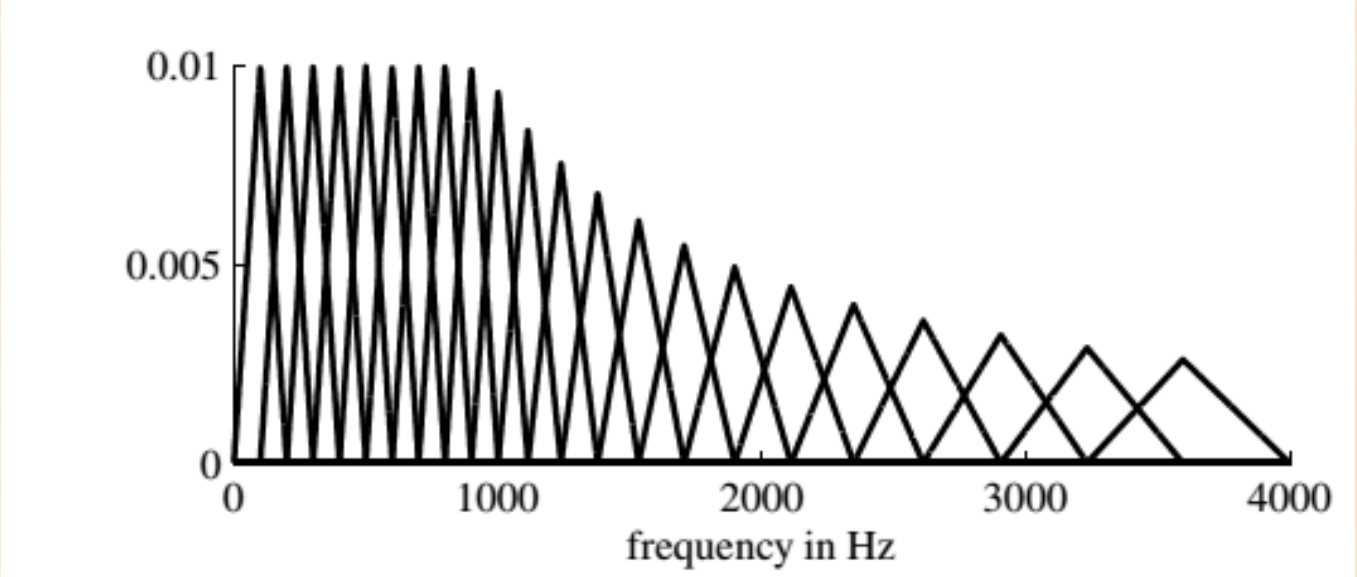
In most implementations, a short-time Fourier analysis is done first, resulting in a DFT $X_{\hat{n}}[k]$ for analysis time $\hat{n}$. Then the DFT values are grouped together in critical bands and weighted by a triangular weighting function

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j(2\pi k/N)n} \tag{5.5a}$$

$$\hat{X}[k] = \log|X[k]| + j \arg\{X[k]\} \tag{5.5b}$$

$$\tilde{\hat{x}}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}[k] e^{j(2\pi k/N)n}. \tag{5.5c}$$

MFCC



the bandwidths are constant for center frequencies below 1 kHz and then increase exponentially up to half the sampling rate of 4 kHz resulting in a total of 22 “filters.”

MFCC

The mel-frequency spectrum at analysis time $\hat{n}$ is defined for r=1,2,...,R as

$$\mathrm{MF}_{\hat{n}}[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r[k] X_{\hat{n}}[k]|^2, \qquad (5.25a)$$

where $V_r[k]$ is the triangular weighting function for the rth filter ranging from DFT index L_r to U_r, where

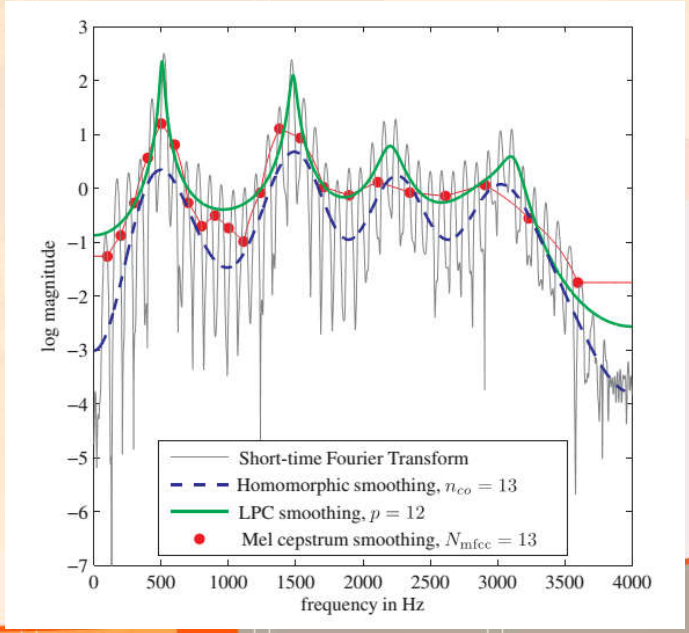
$$A_r = \sum_{k=L_r}^{U_r} |V_r[k]|^2 \qquad (5.25b)$$

is a normalizing factor for therth mel-filter.

MFCC

For each frame, a discrete cosine transform of the log of the magnitude of the filter outputs is computed to form the function $\mathrm{mfcc}_{\hat{n}}[m]$

$$\mathrm{mfcc}_{\hat{n}}[m] = \frac{1}{R}\sum_{r=1}^{R} \log\left(\mathrm{MF}_{\hat{n}}[r]\right) \cos\left[\frac{2\pi}{R}\left(r+\frac{1}{2}\right)m\right]. \qquad (5.26)$$



MFCC

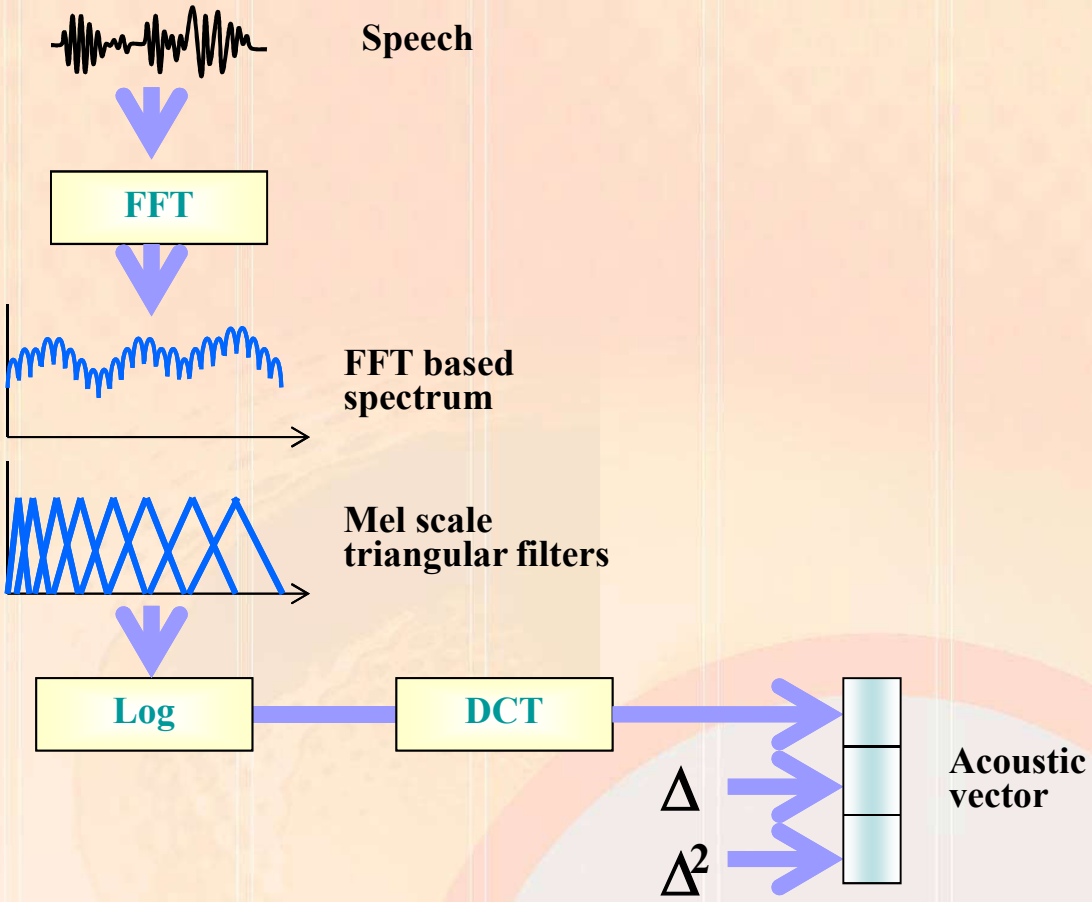
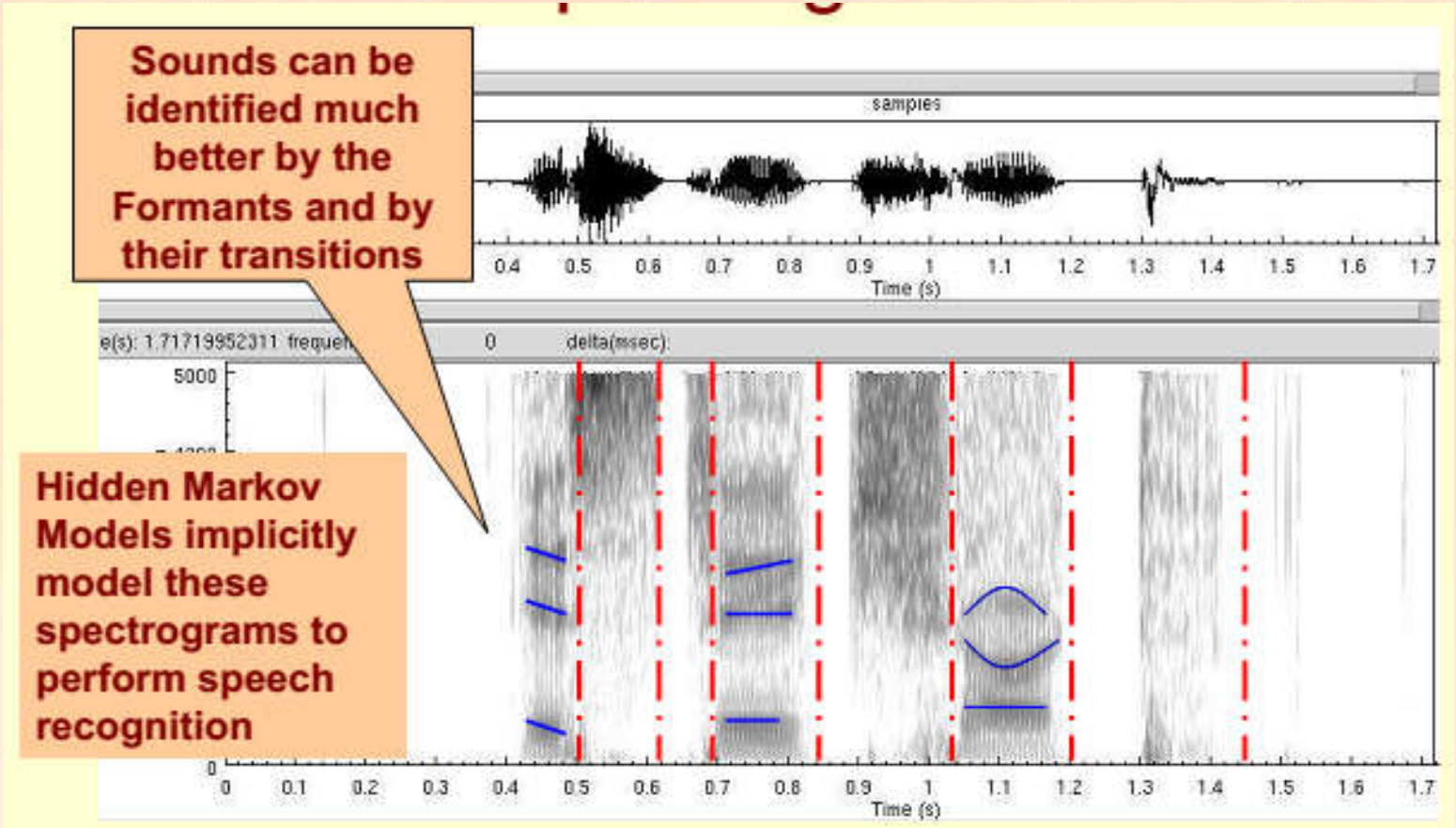


Fig.4. MFCC feature extraction

Problem Statement



参考文献

1.吴朝晖，杨莹春，说话人识别模型与方法，清华大学出版社，2009, 2

2.杨莹春，陈华，吴飞，视音频信号处理，浙江大学出版社，待出版

3. Roger Jang (張智星)

Audio Signal Processing and Recognition (音訊處理與辨識)
http://neural.cs.nthu.edu.tw/jang/books/audioSignalProcessing/index.asp

课后任务

- 阅读文献
  - L. R. Rabiner and R. W. Schafer, Introduction to Digital Speech Processing

Ch4_Short-Time Analysis of Speech (4.1-.5,4.8)

Ch3_Hearing and Auditory Perception(3.1-3.4)

Ch5_Homomorphic Speech Analysis(5.1,5.2,5.3.1,5.4,5.6.3,5.7)