数字语音处理II
Digital Speech Processing

https://courses.zju.edu.cn/course/join/8P84SXUIRG2

访问码: 8P84SXUIRG2

频域分析
Frequency Analysis

杨莹春

yyc@zju.edu.cn

浙江大学教7-506

2021年10月8日

教学安排

讲授内容

(9月17日）秋1：课程简介＋语音技术引言
(9月24日）秋2：语音时域分析
(10月8日）秋4：语音频域分析&语音识别
(10月15日）秋5：说话人识别、语音编码及合成
(12月17日）冬6：复习及项目成果展示（加实验课）

实验内容

1. PRAAT 语音分析 （9月24日）秋2
2. VOICEBOX说话人识别 （10月16日）补秋6

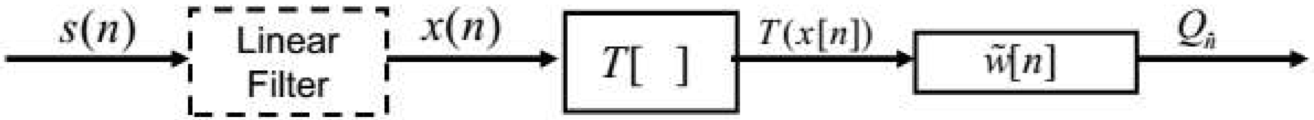
3、项目展示（ 12月17日）冬6

考试：2022年1月

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:

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

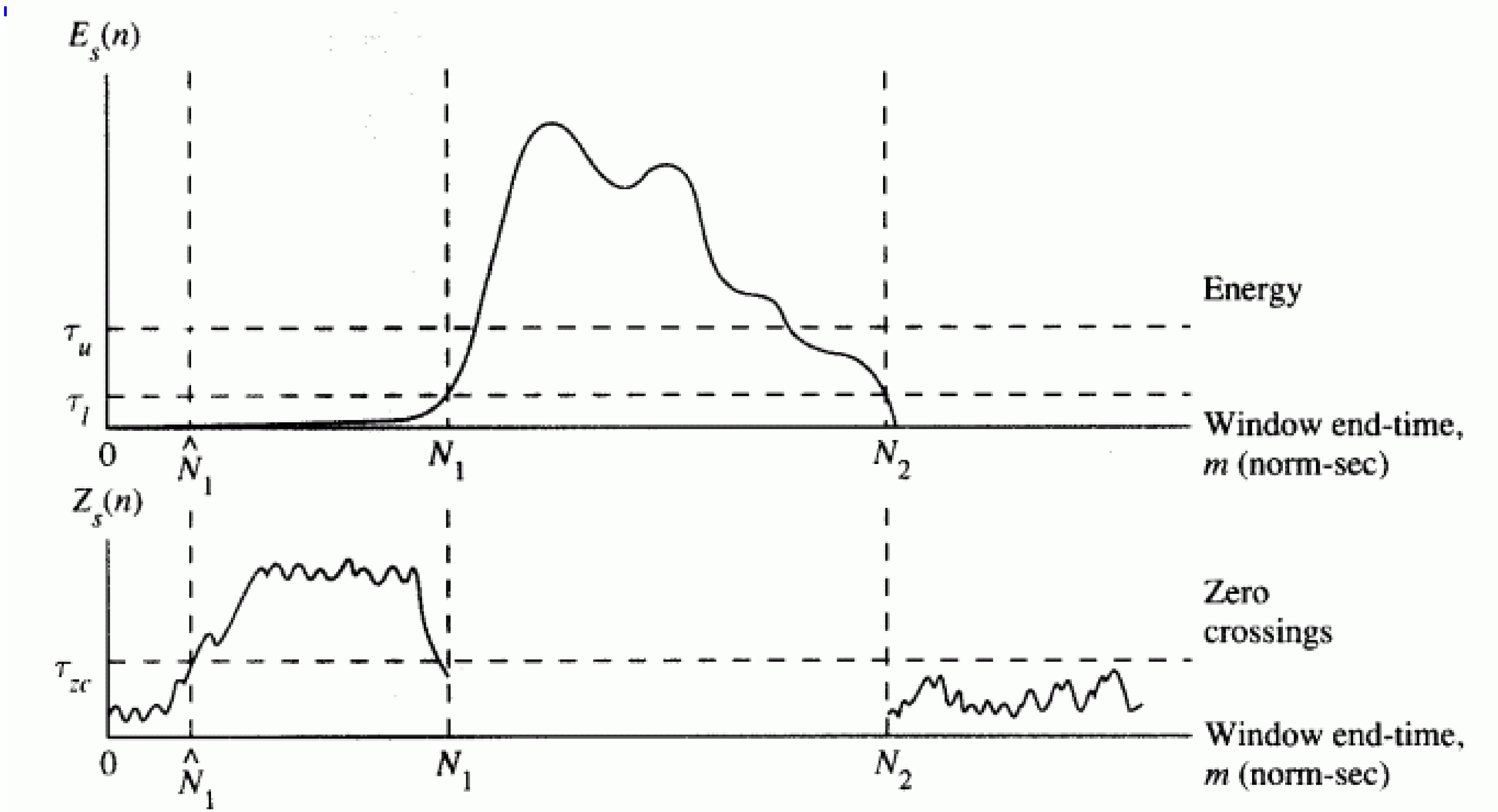
3. Zero Crossing Rate:

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

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

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

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

<u>Algorithm</u> for endpoint detection:

1. compute mean and σ of E_n and Z_n for first 100 msec of signal (<u>assuming no speech in this interval</u>).
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.

auto-correlation function (ACF)

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

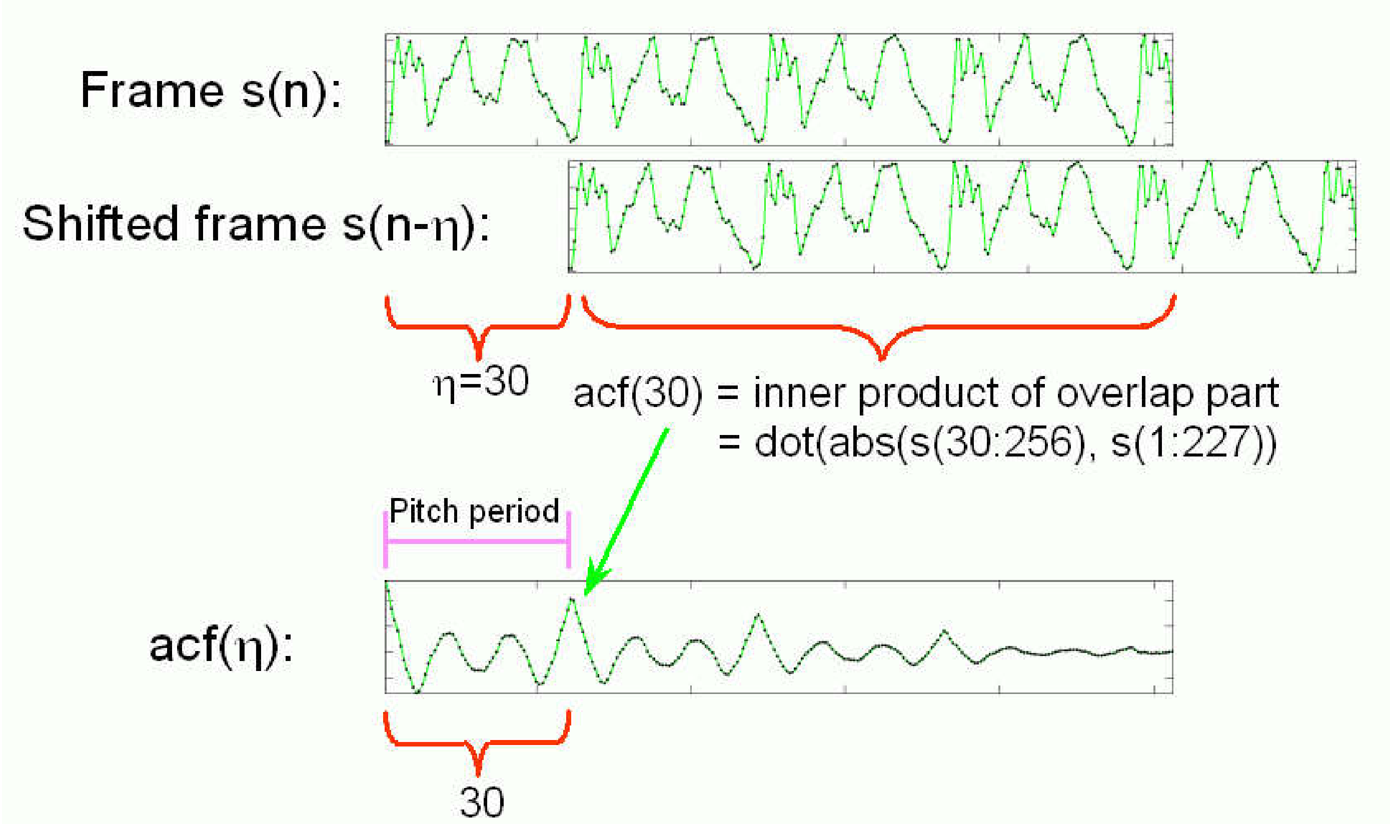
and its delayed version via the auto-correlation function:

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

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.

语音分析技术

- 语音时域分析
- 语音频域分析

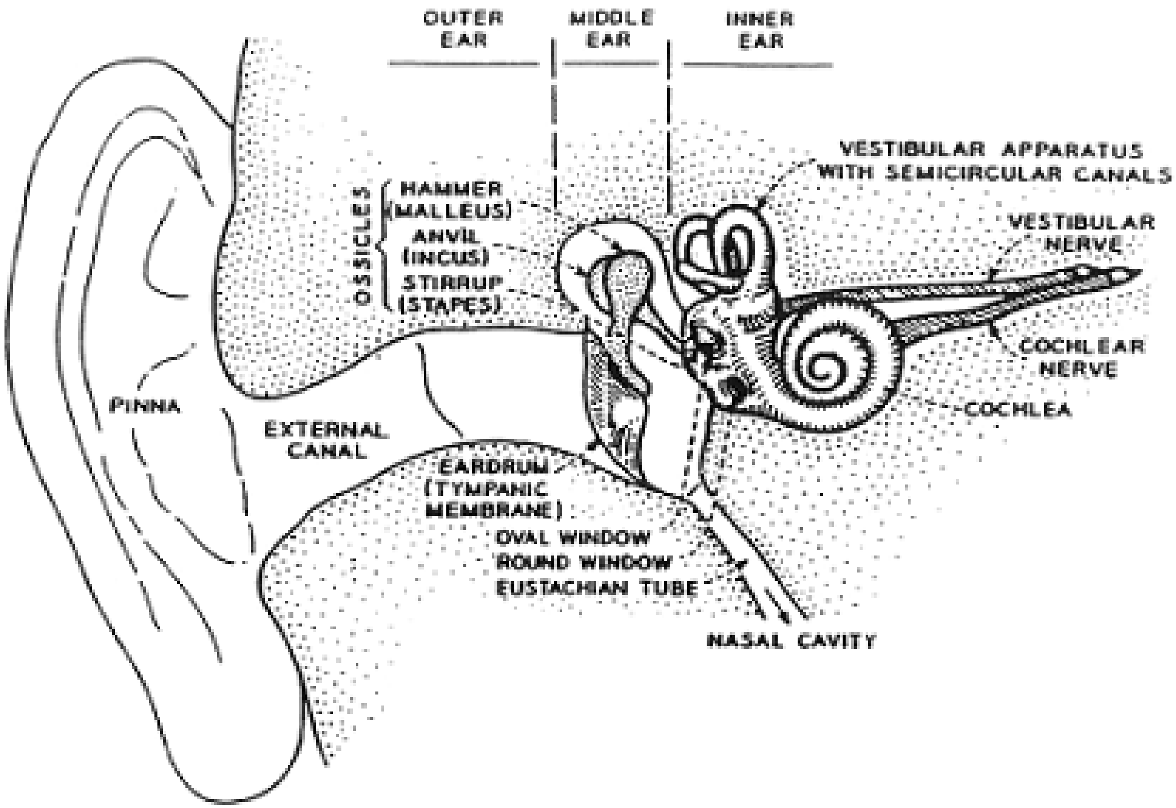
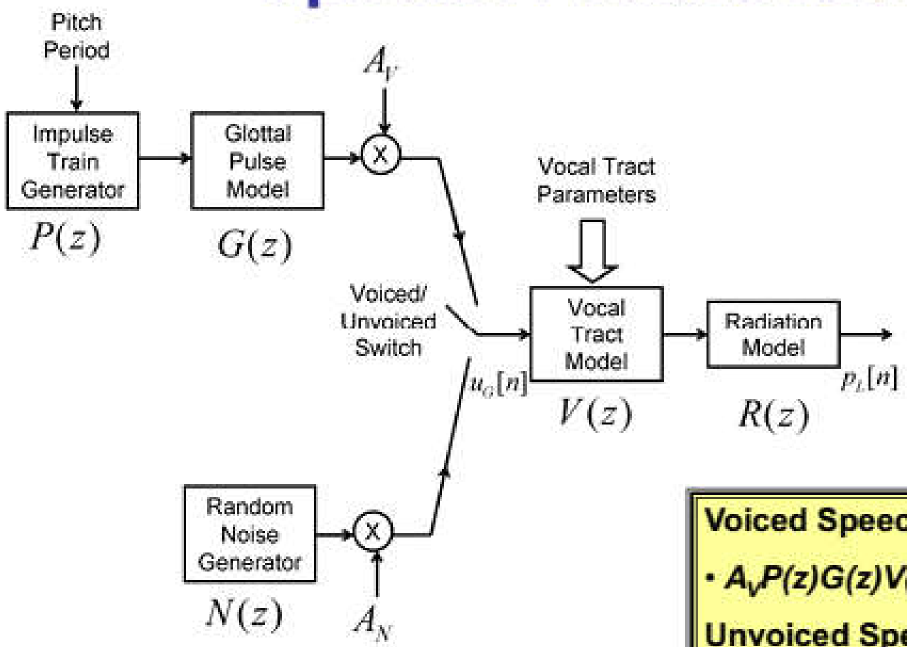


Fig. 3.1 Schematic view of the human ear (inner and middle structures enlarged). (After Flanagan [34].)

General Discrete-Time Model of Speech Production



Voiced Speech:

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

Unvoiced Speech:

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

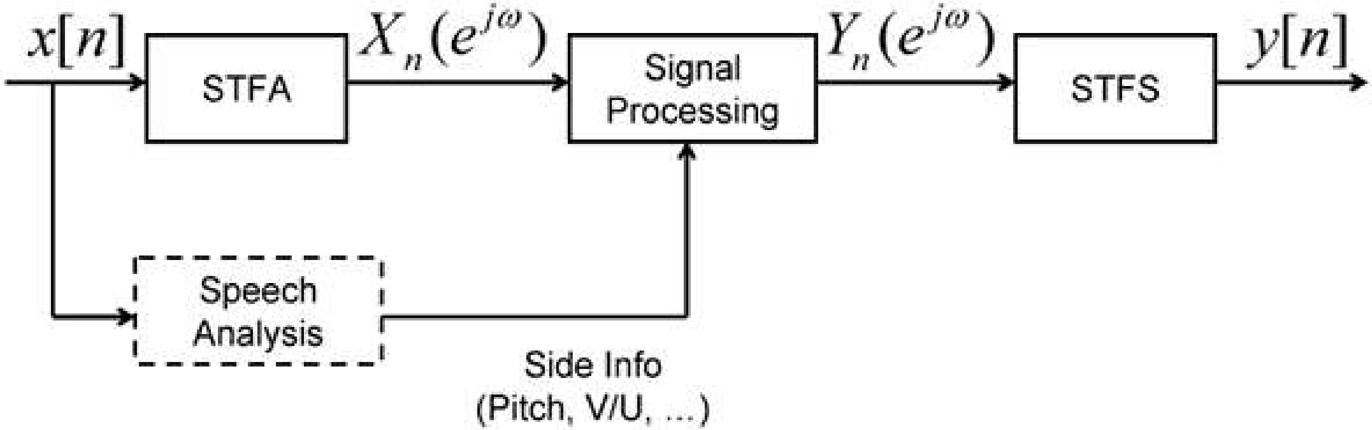
Short-Time Fourier Analysis

- 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

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)

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)

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 $X(e^{j\omega})$

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:

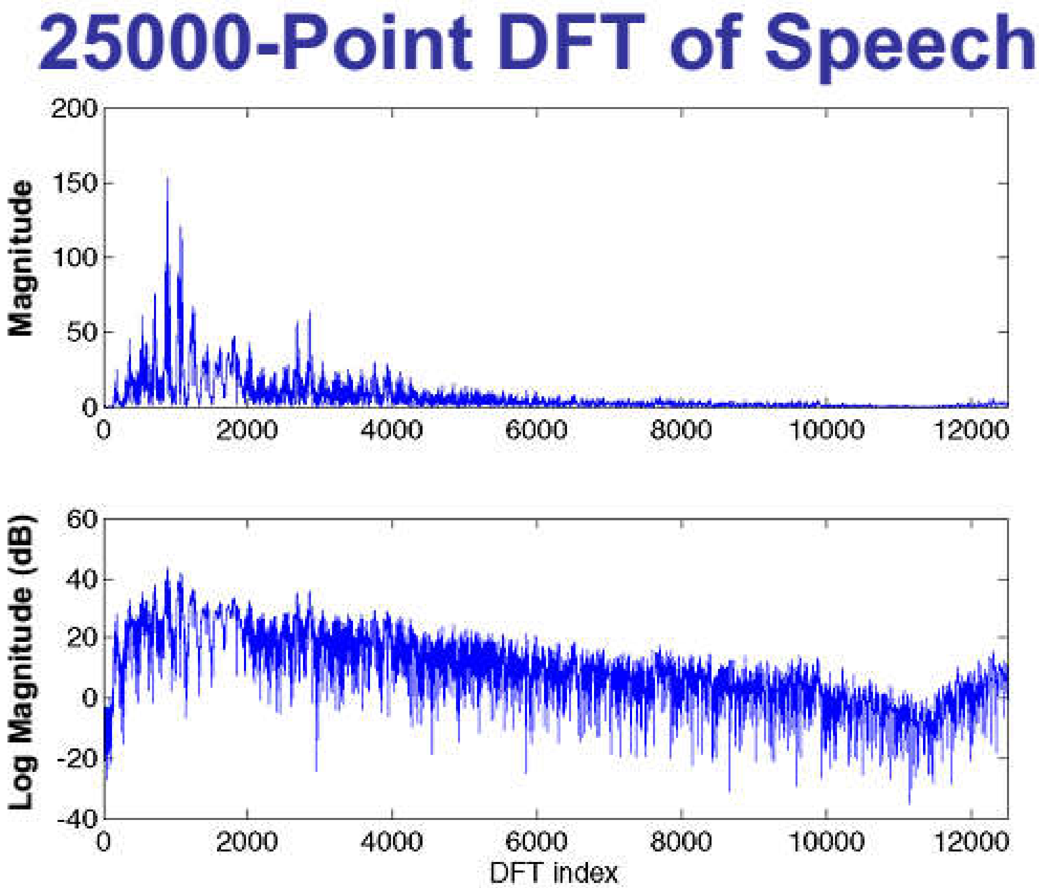
$$X(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)e^{-j\omega m} \quad (DTFT)$$

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

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

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

Short-Time Fourier Analysis短时傅里叶分析

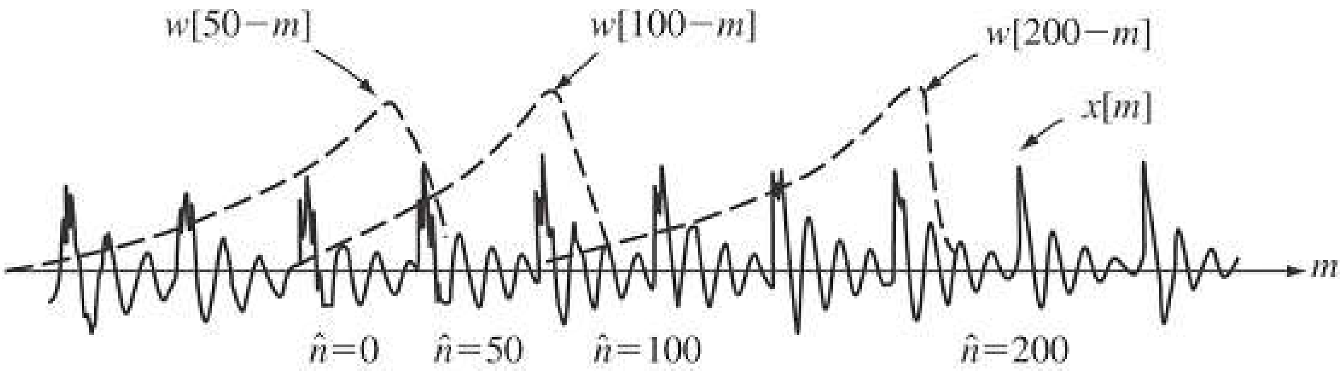


Definition of STFT

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

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 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

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

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

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

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 $X_{\hat{n}}(e^{j2\pi\hat{F}T})$

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

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

Short-Time Fourier Analysis短时傅里叶分析

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

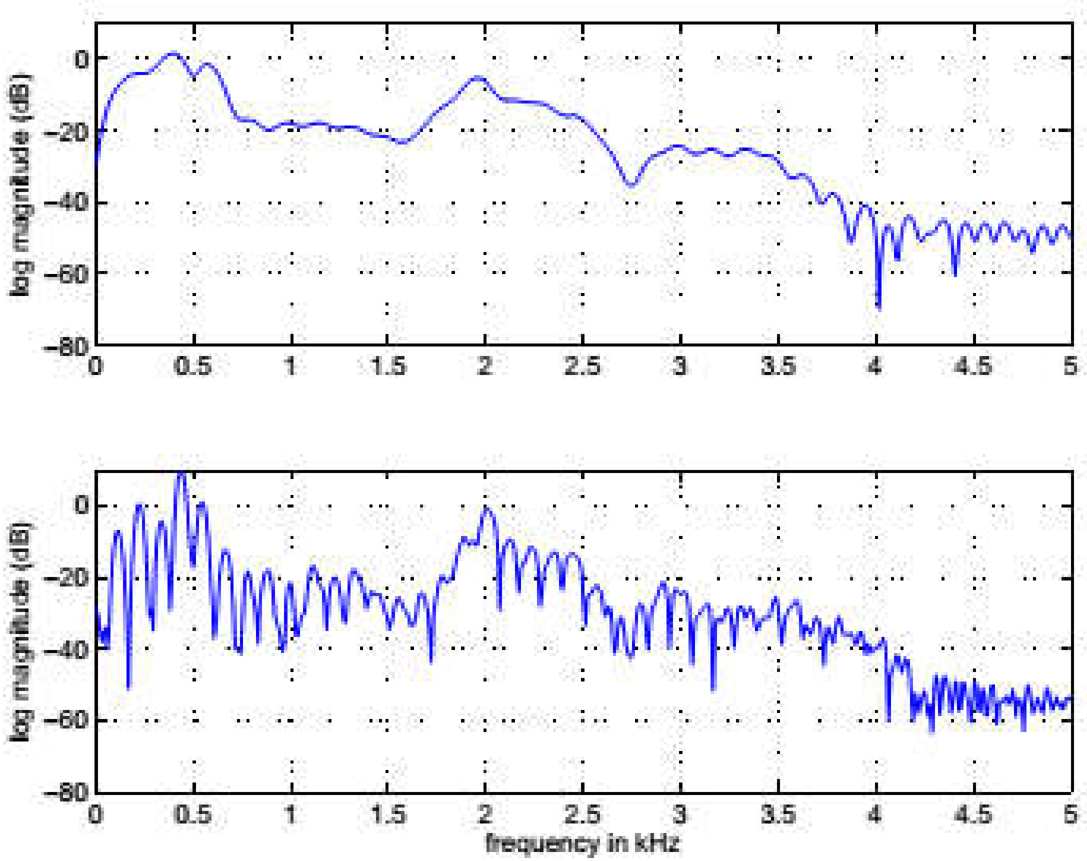


Fig. 4.7 Short-time spectrum at time 430 ms (dark vertical line in Figure 4.6) with Hamming window of length $M = 101$ in upper plot and $M = 401$ in lower plot.

cience

Speech spectrogram

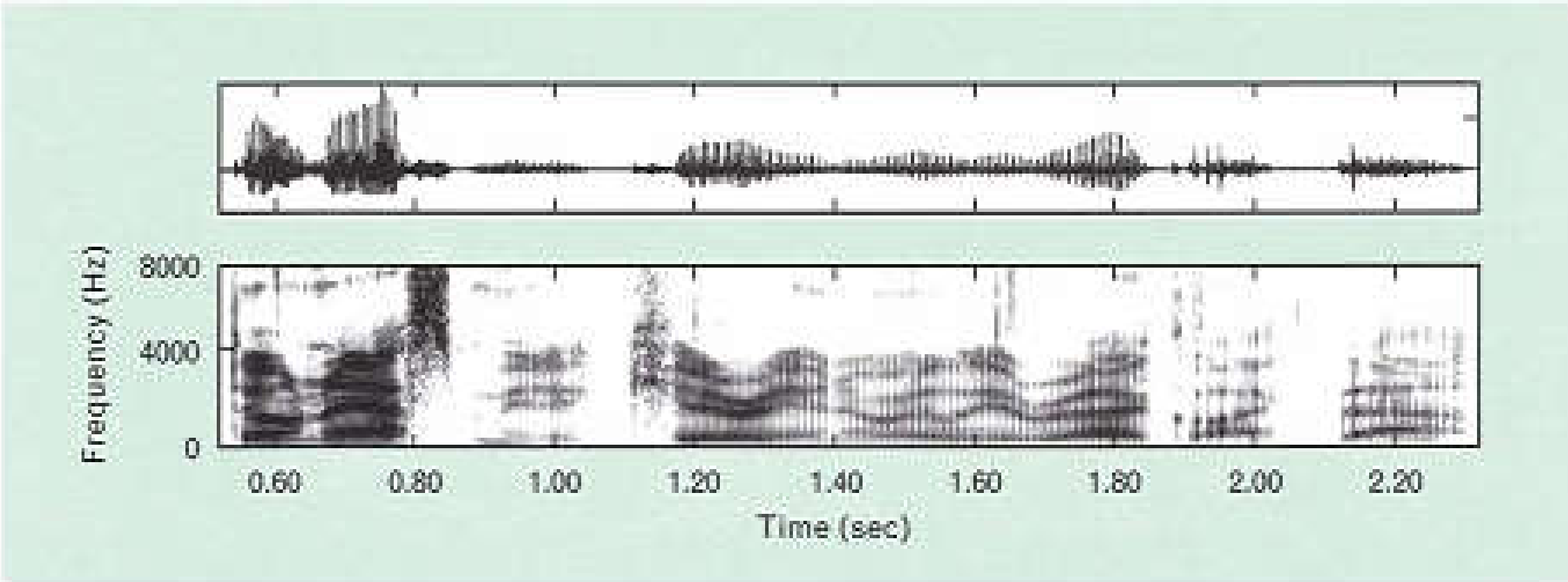
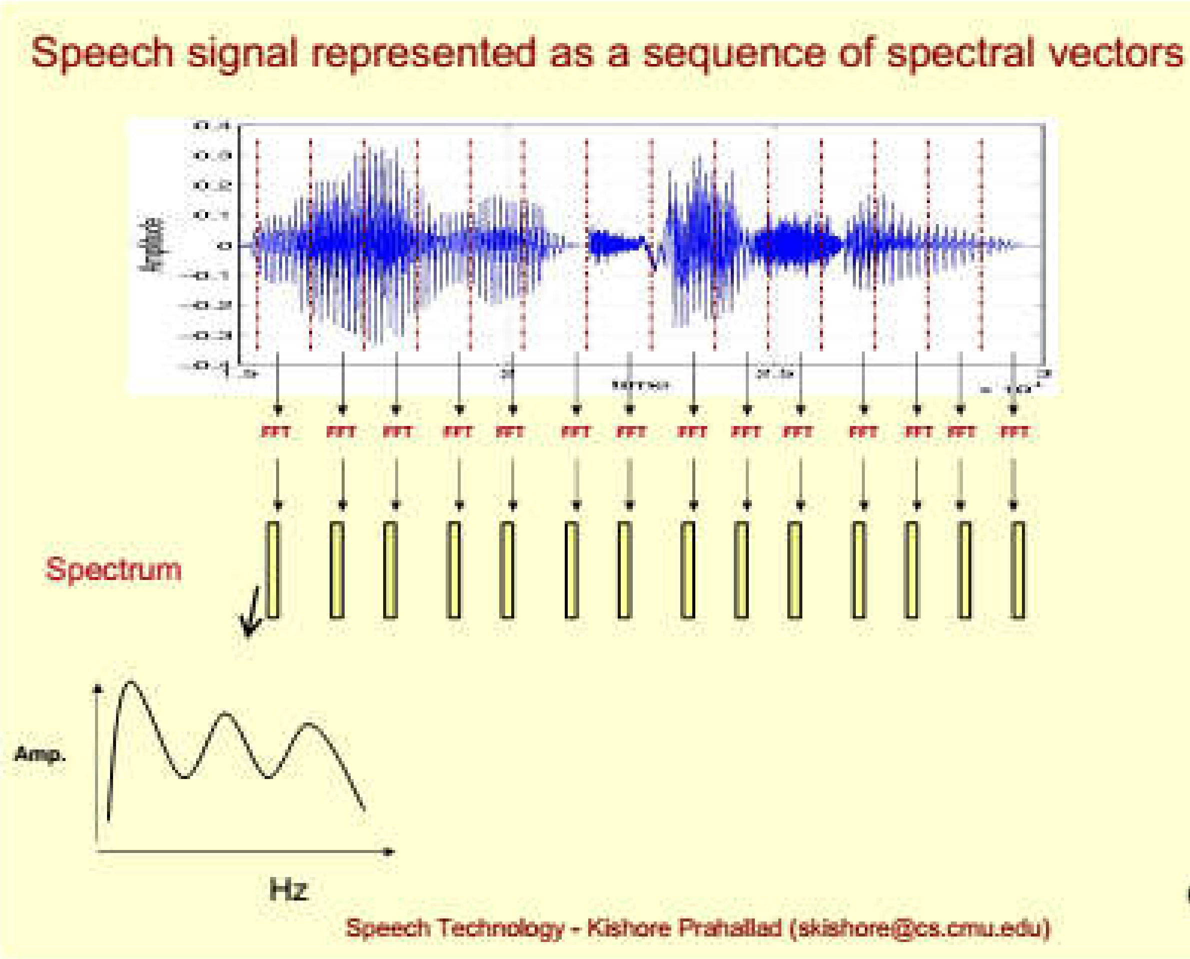
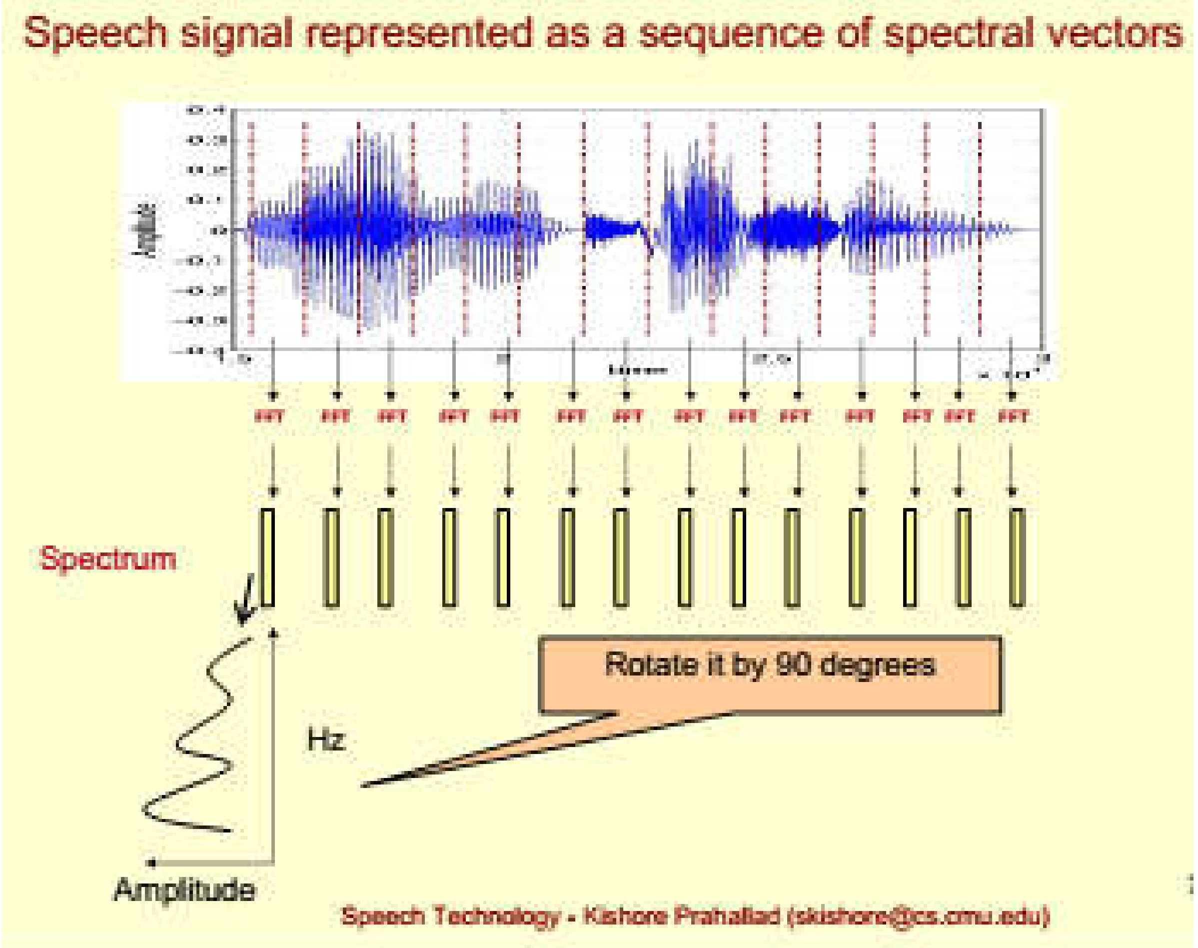


FIGURE 3. Digitally sampled speech waveform of a spoken sentence (above) and corresponding spectrogram (below) showing the dynamic nature of the formants as the vocal tract continuously changes shape. The sentence spoken was "Don't ask me to carry an oily rag like that."

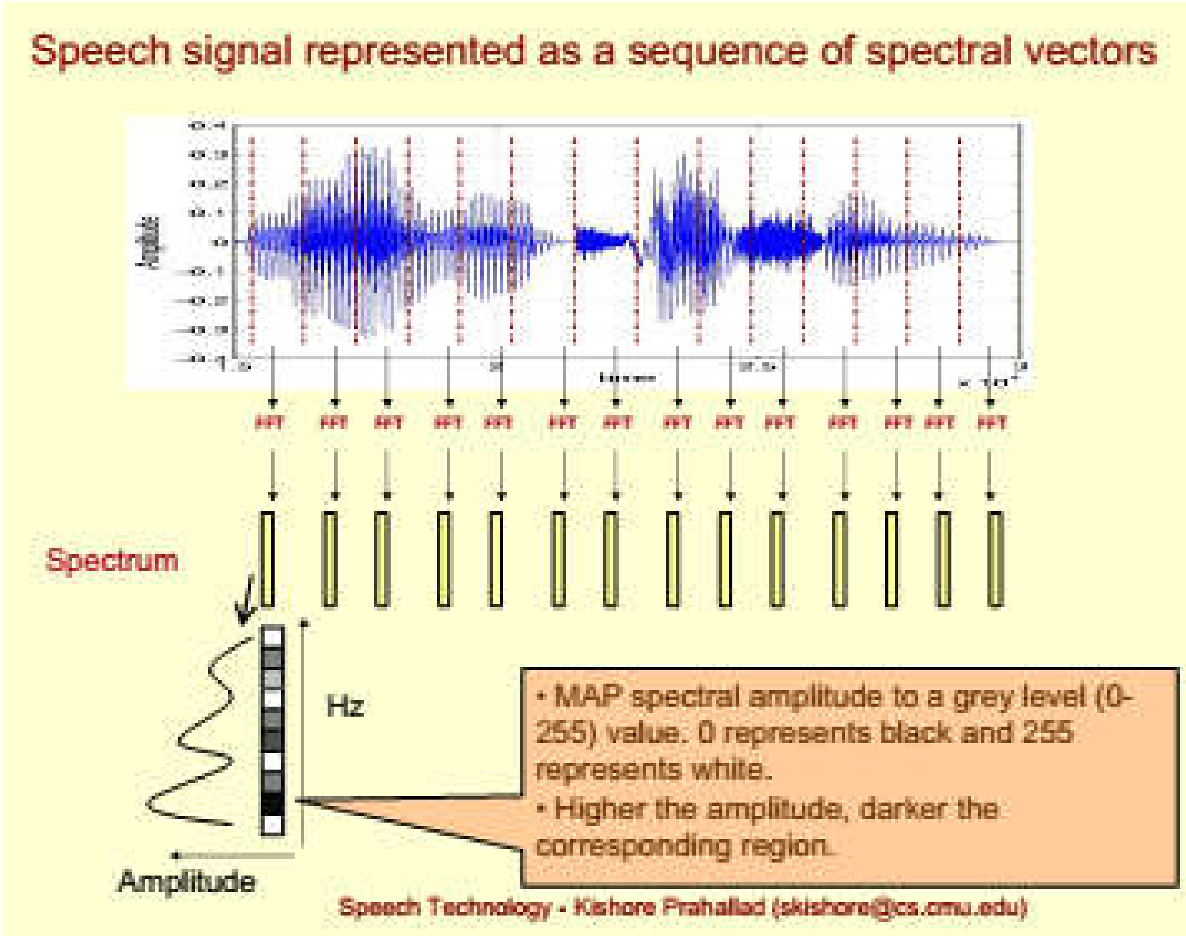
Speech spectrogram



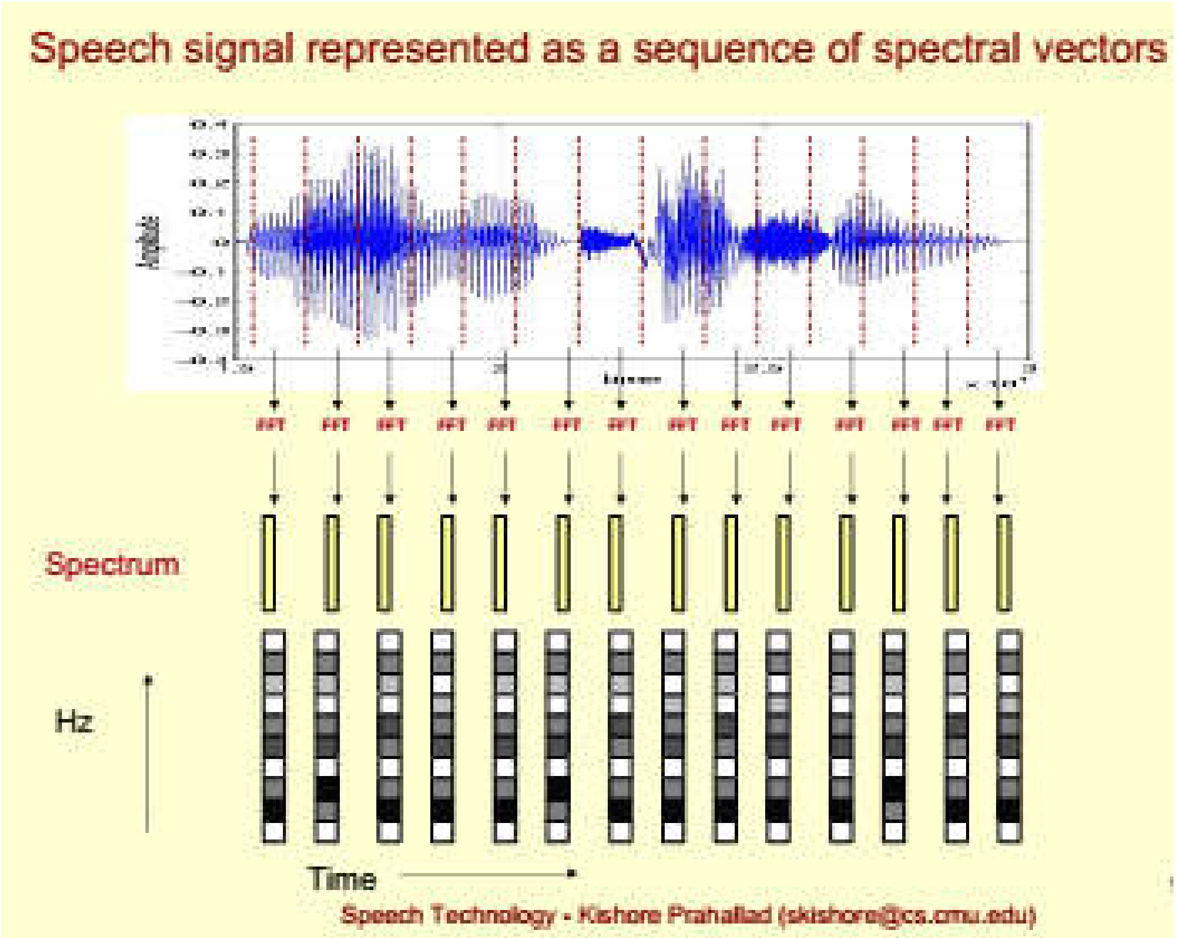
Speech spectrogram



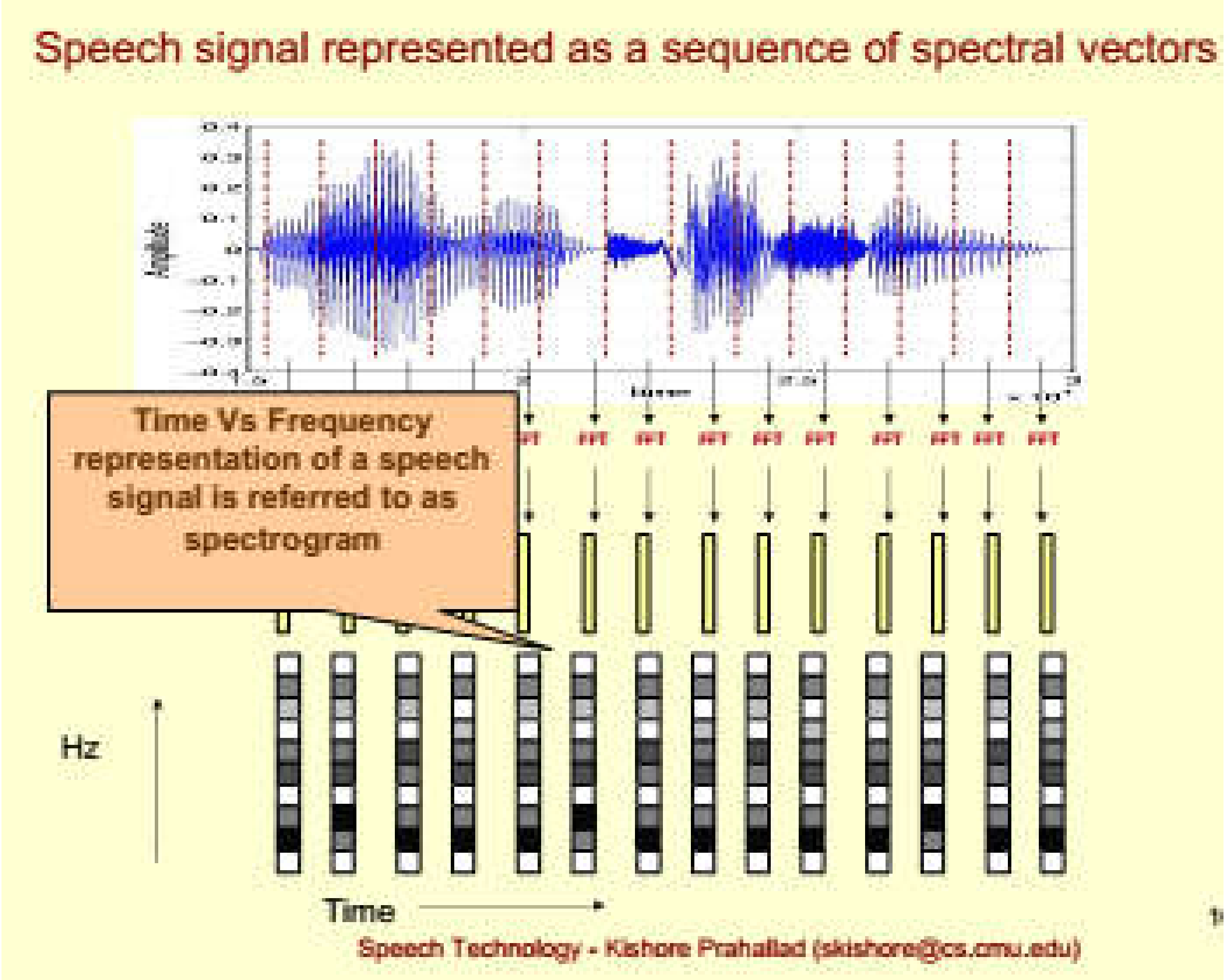
Speech spectrogram



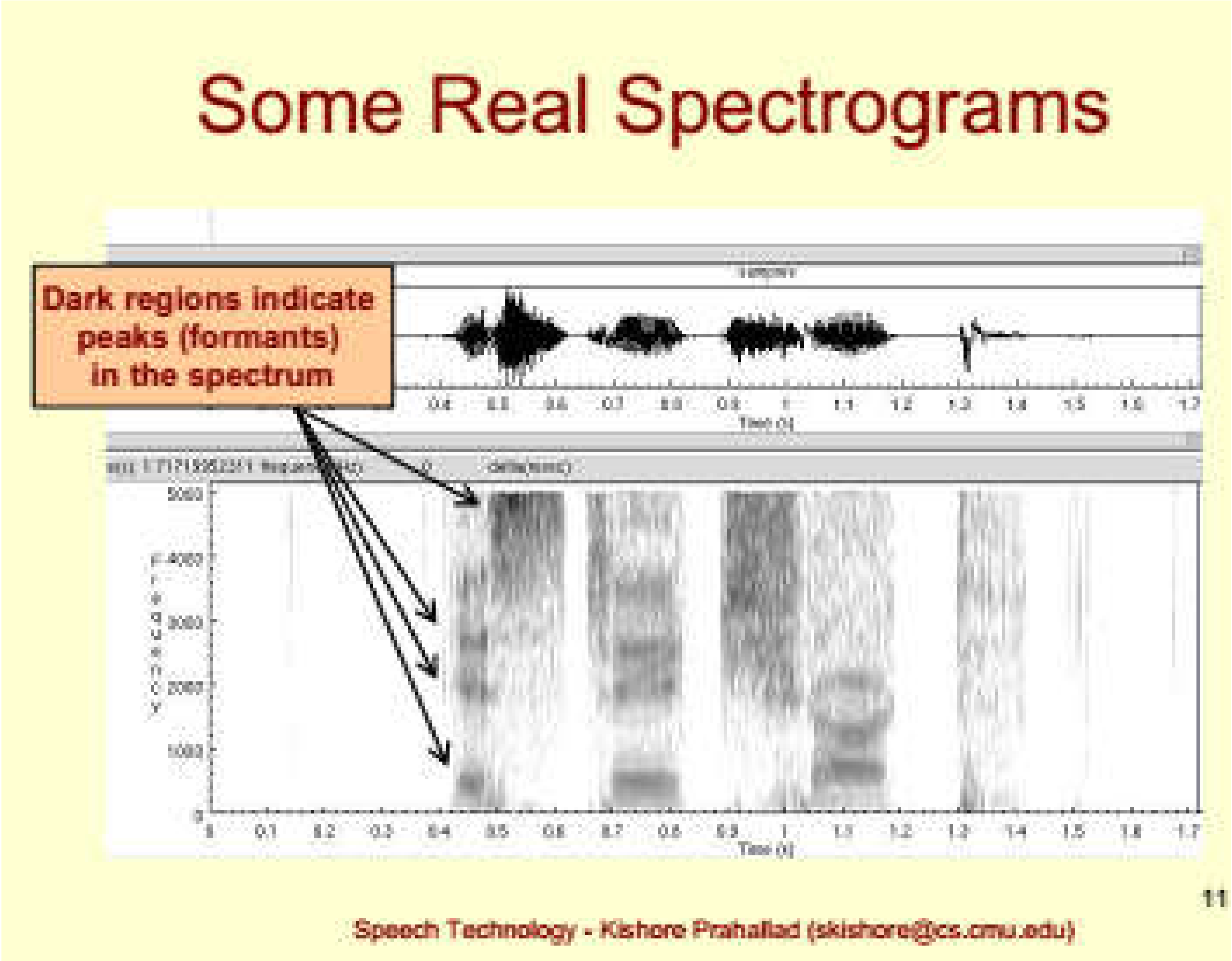
Problem Statement



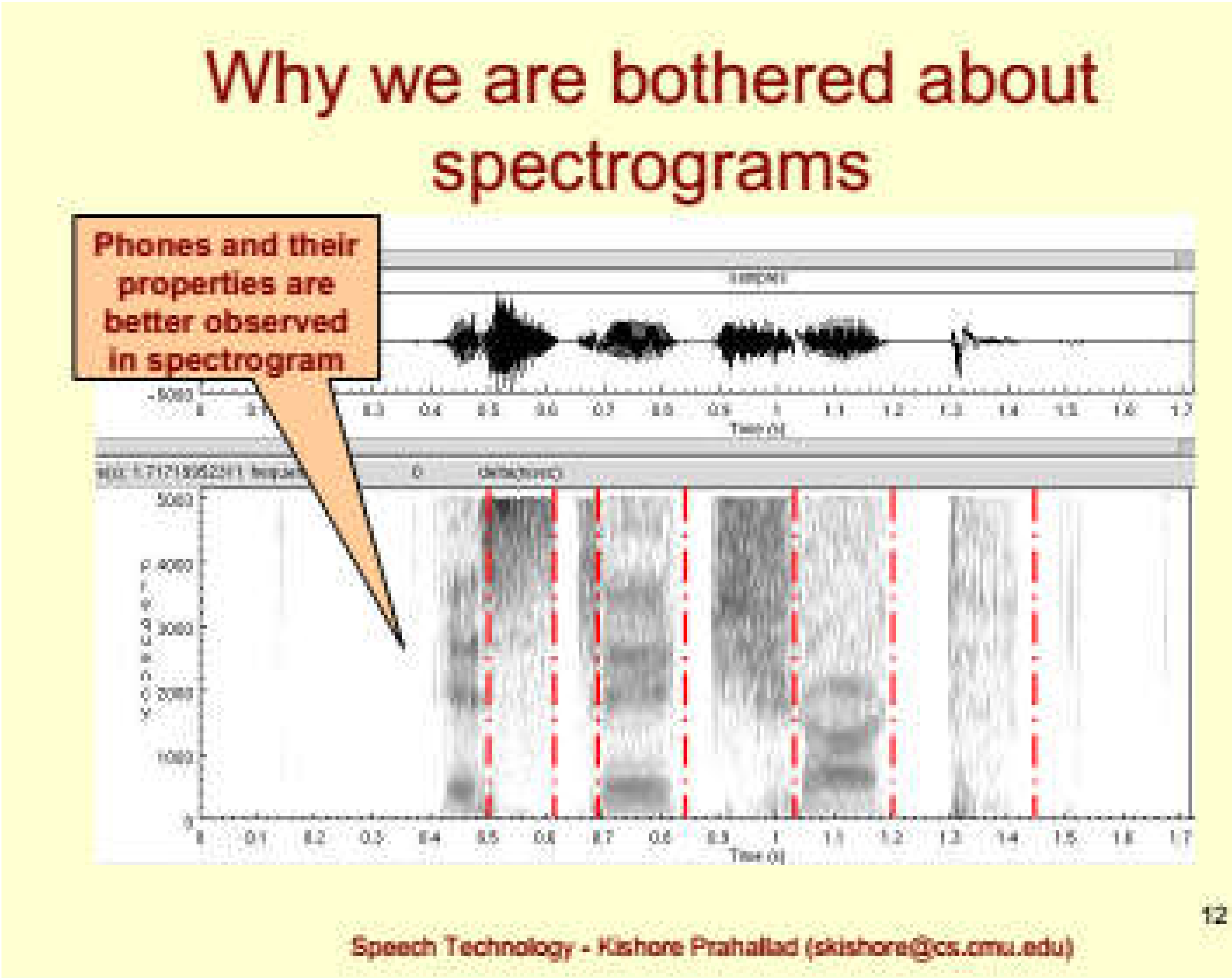
Speech spectrogram



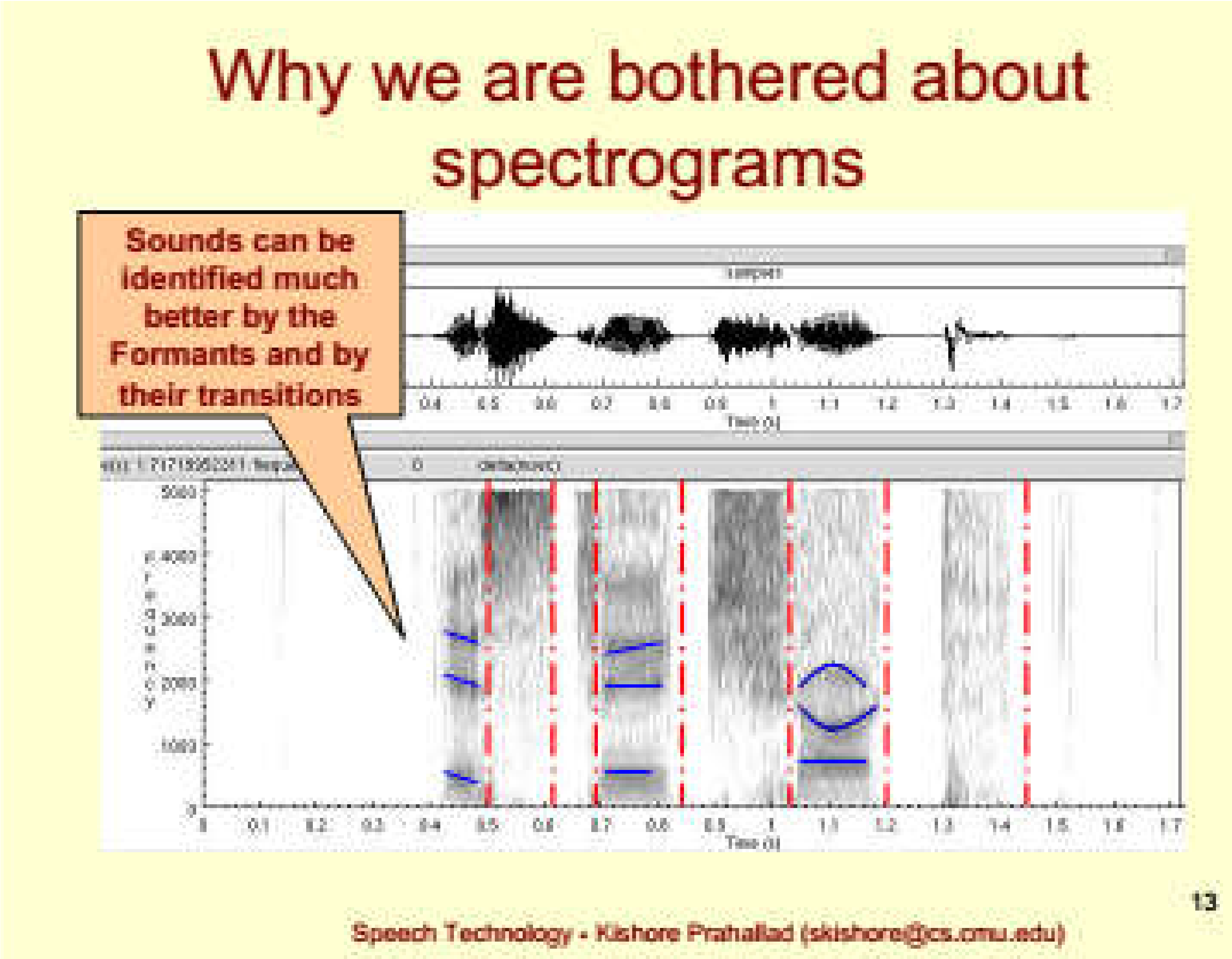
Problem Statement



Speech spectrogram

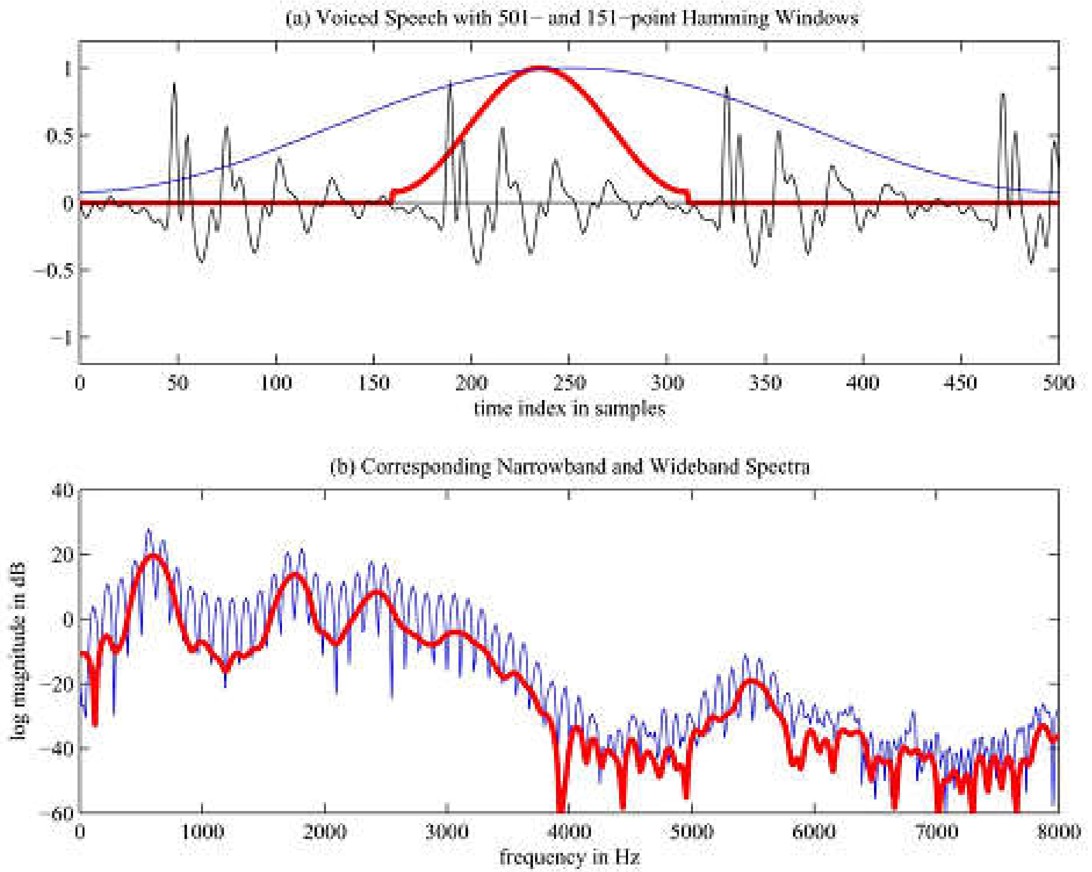


Speech spectrogram



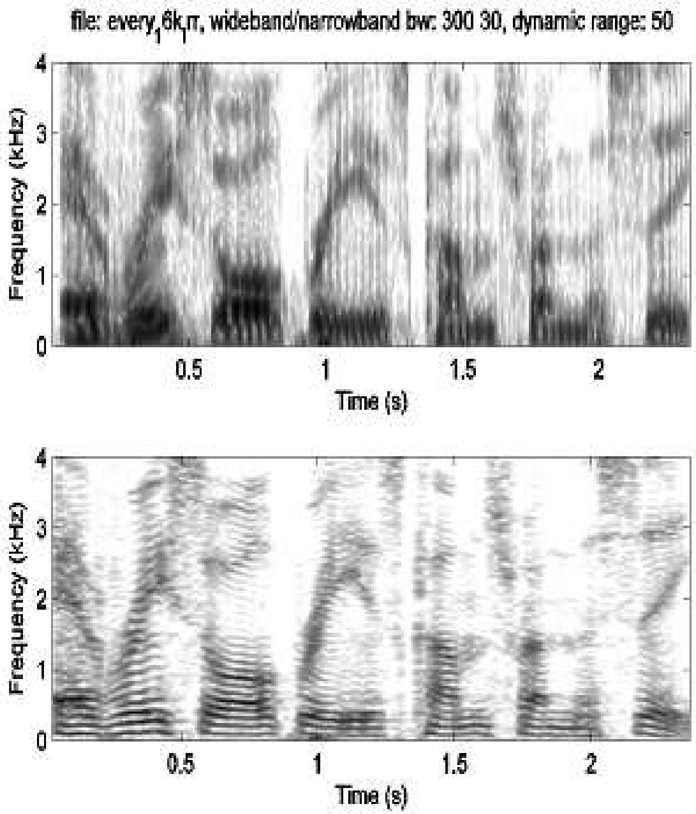
Short-Time Fourier Analysis短时傅里叶分析

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

Cepstrum analysis

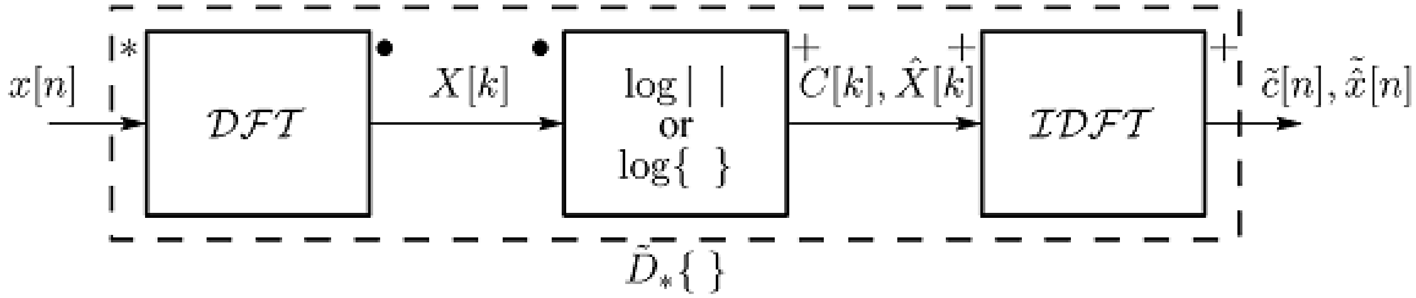
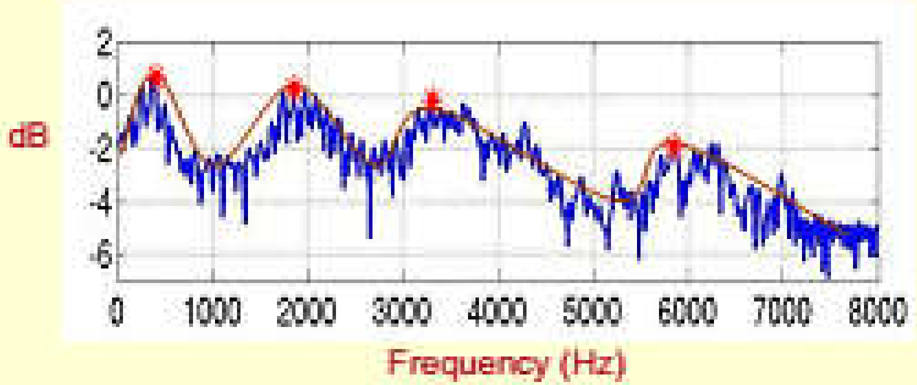


Fig. 5.3 Computing the cepstrum or complex cepstrum using the DFT.

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

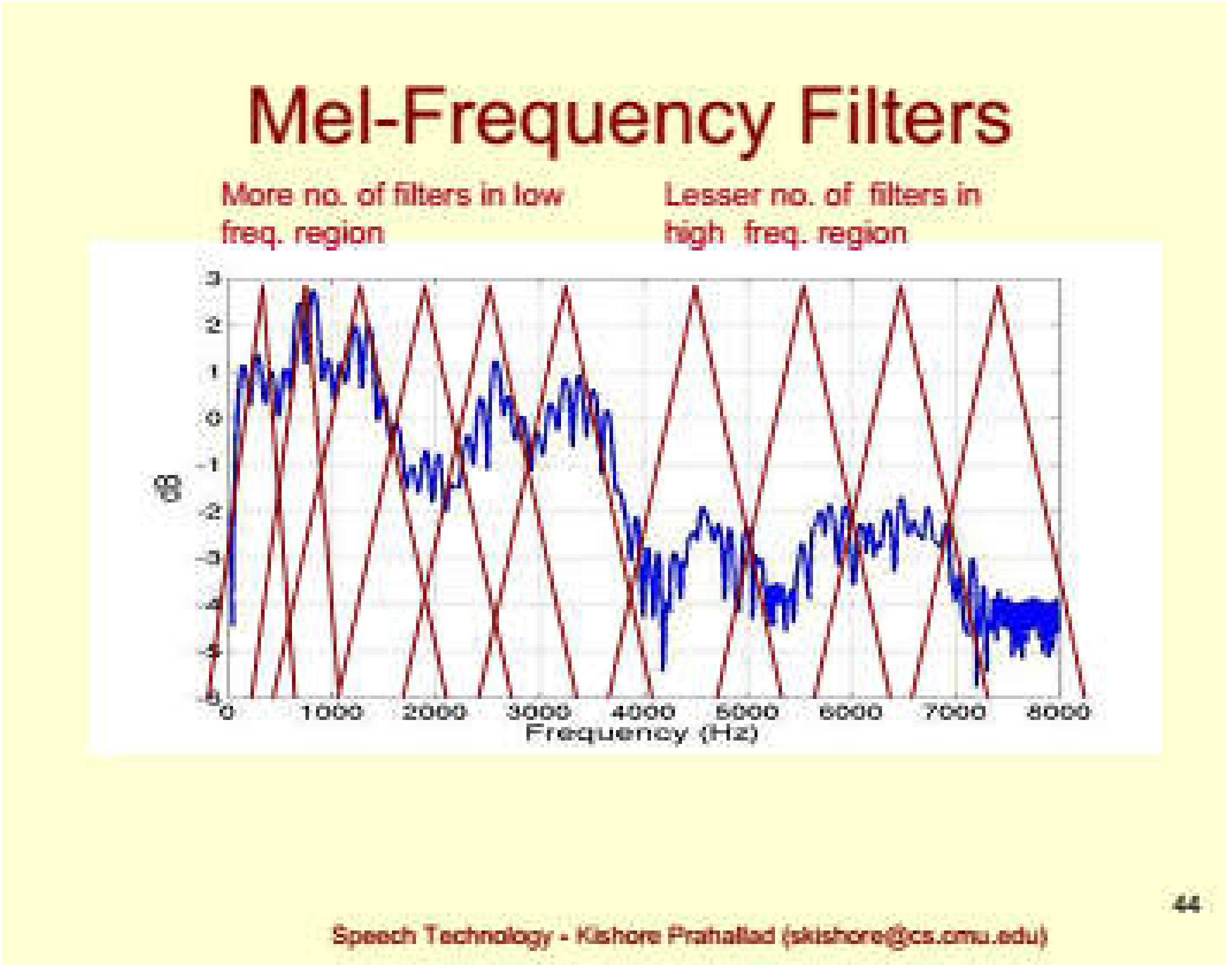
Mel-Frequency Analysis

- 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

42

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

MFCC


Mel-Frequency Filters
More no. of filters in low freq. region
Lesser no. of filters in high freq. region
dB
0
1000
2000
3000
4000
5000
6000
7000
8000
Frequency (Hz)
44
Speech Technology - Kishore Prahallad (skishore@cs.cmu.edu)

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

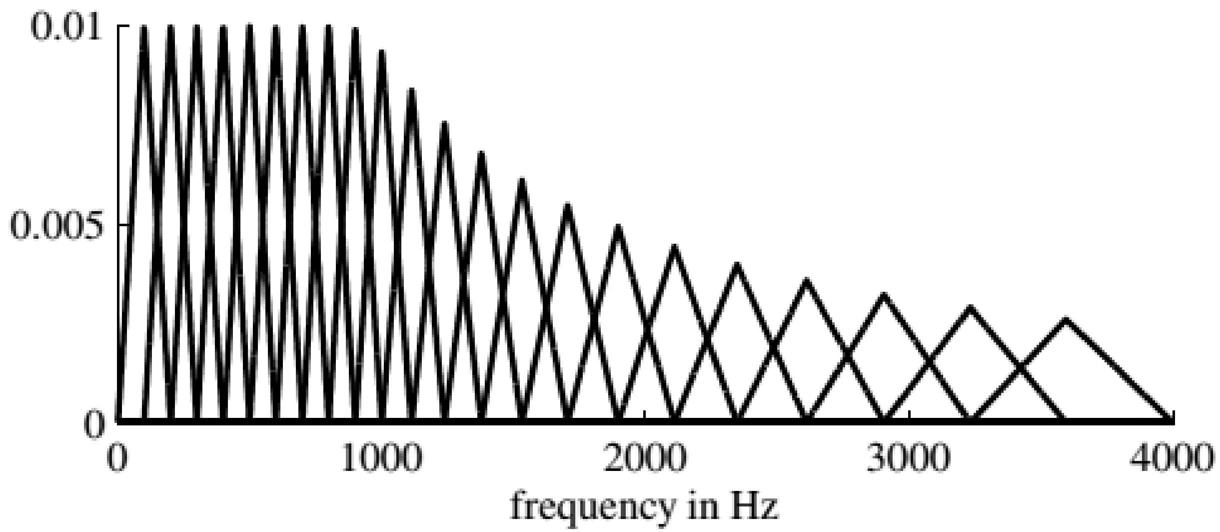
In most implementations, a short-time Fourier analysis is done first, resulting in a DFT $X_{\hat{n}}[k]$ for analysis time $\hat{n}$. Then the DFT values are grouped together in critical bands and weighted by a triangular weighting function

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

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

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

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
$\hat{n}$ me
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

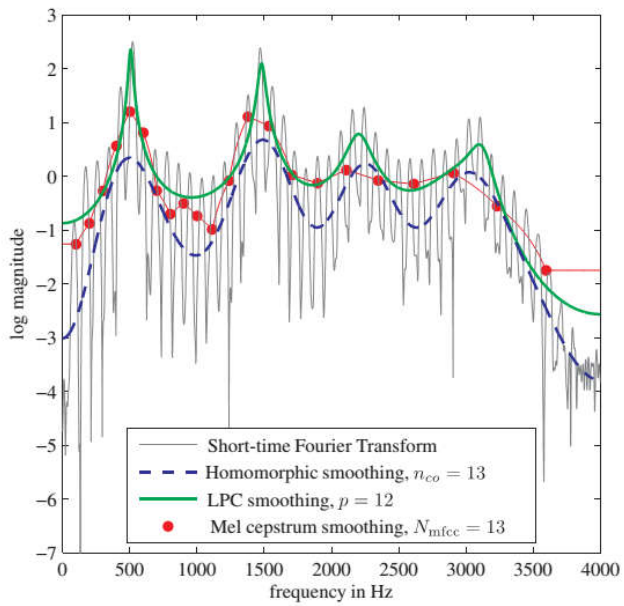
$$A_r = \sum_{k=L_r}^{U_r} |V_r[k]|^2 \qquad (5.25b)$$

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 $\mathrm{mfcc}_{\hat{n}}[m]$n

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



MFCC

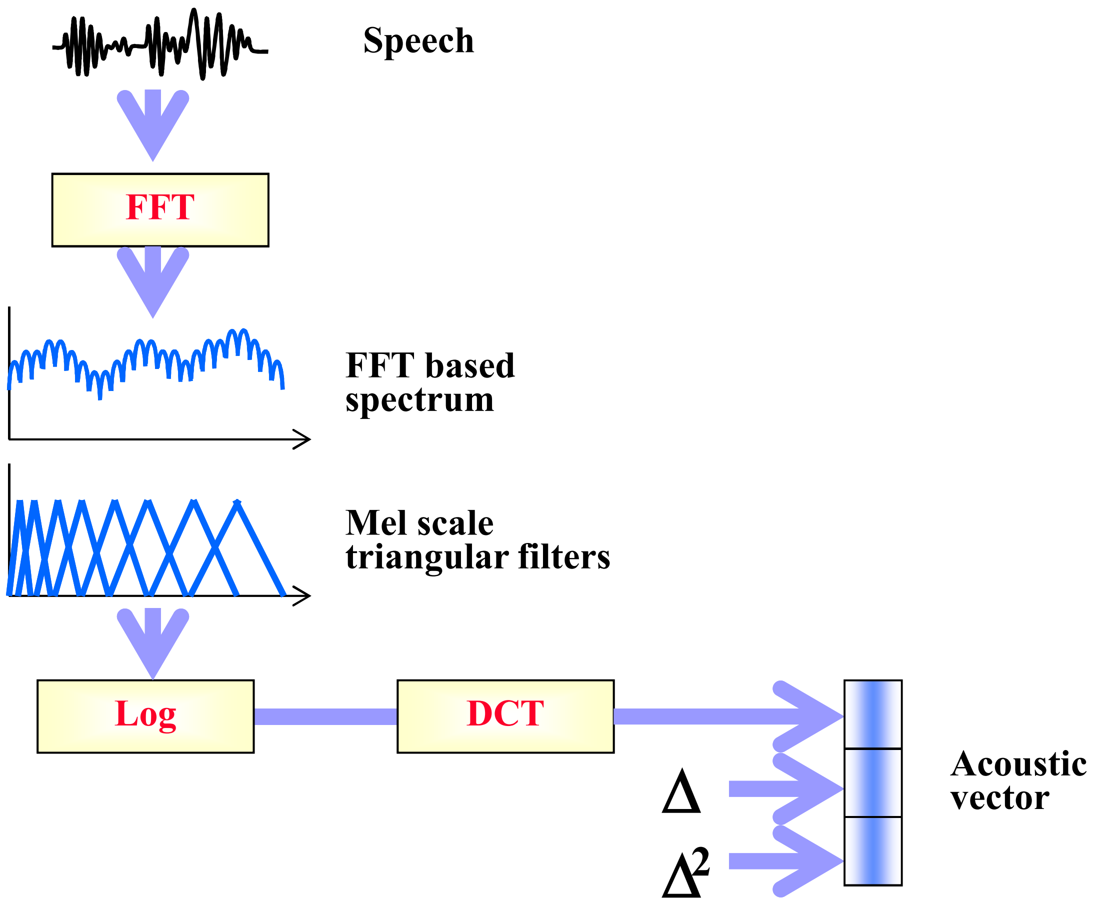
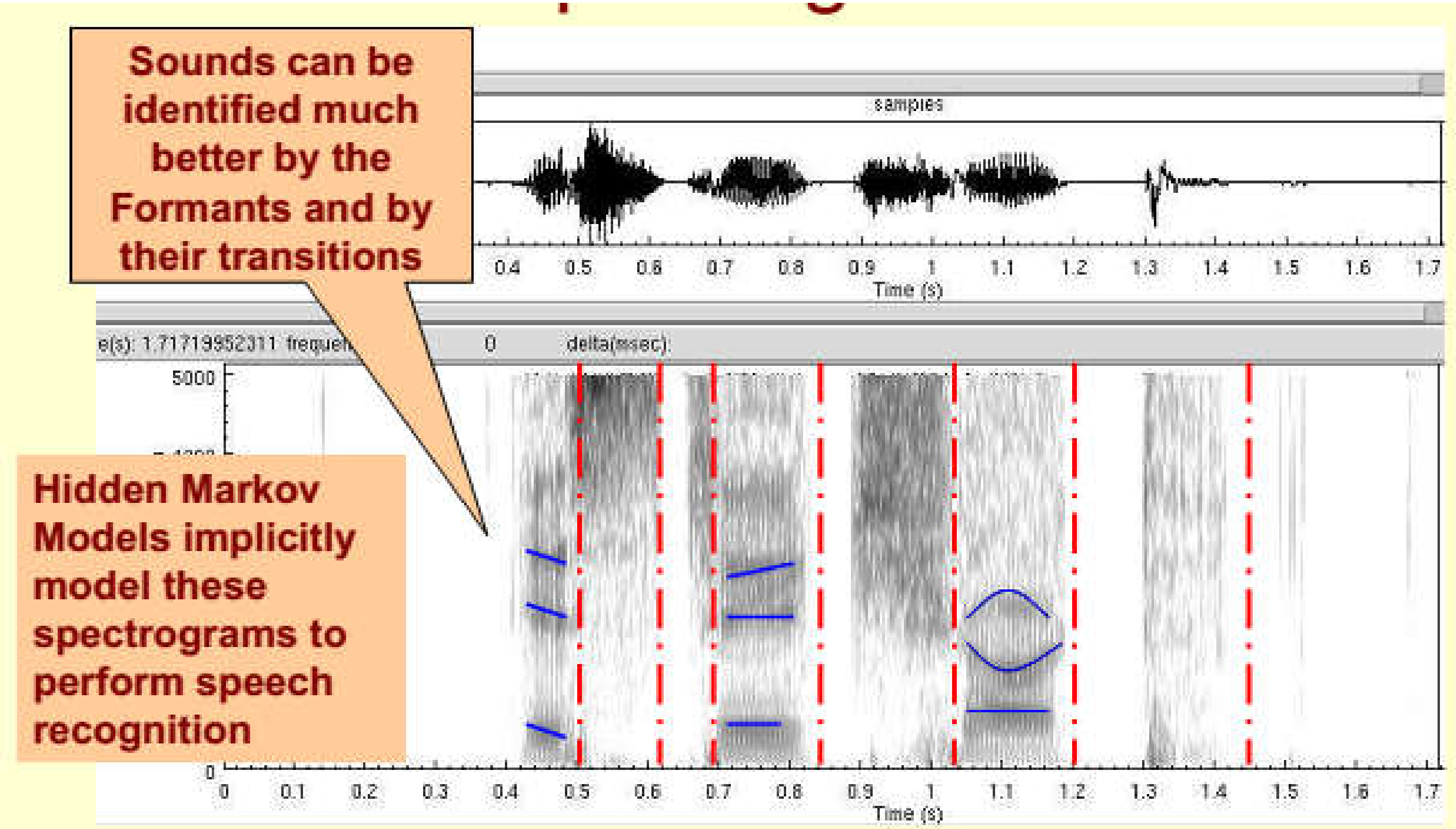


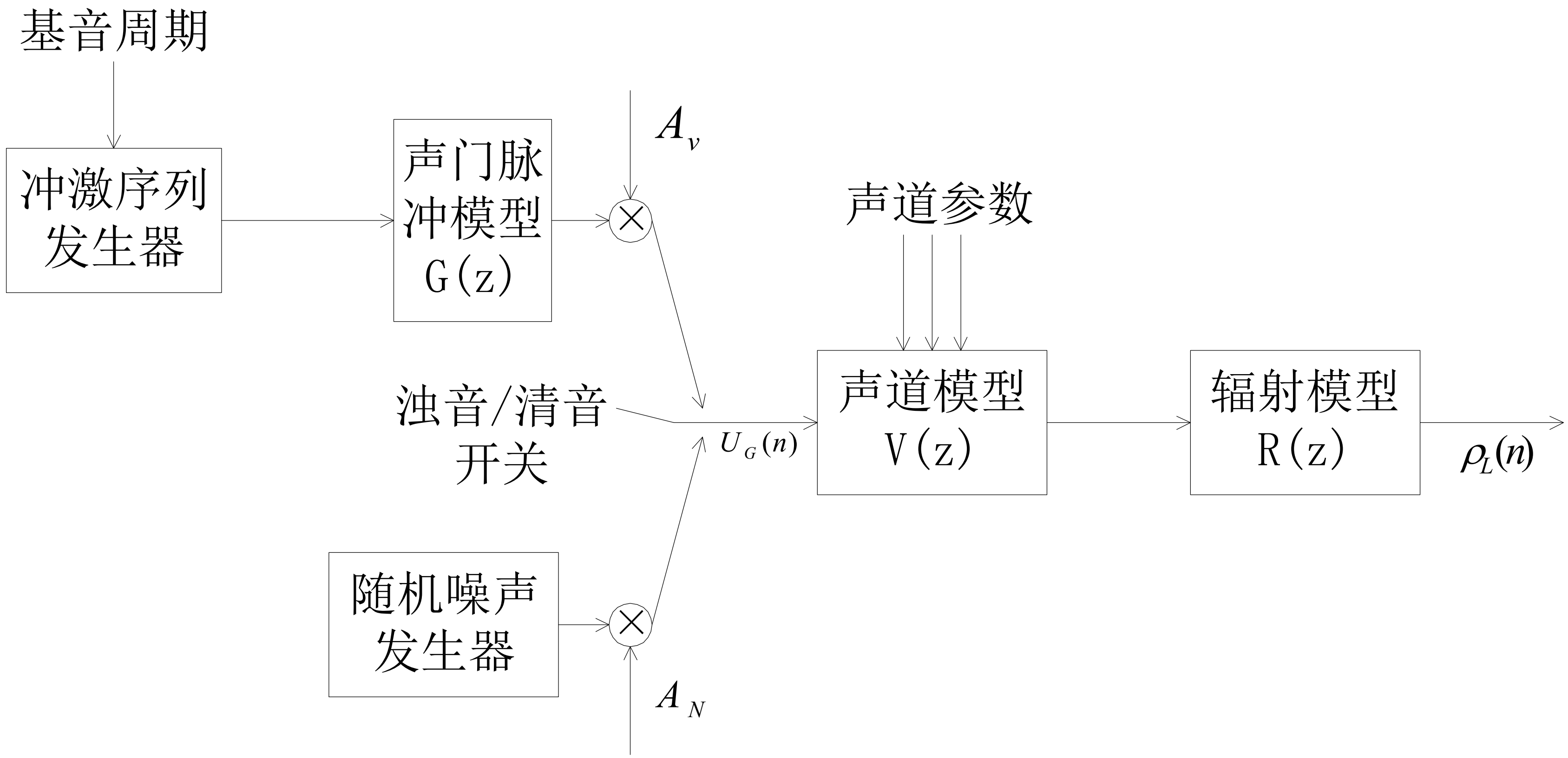
Fig.4. MFCC feature extraction

Problem Statement

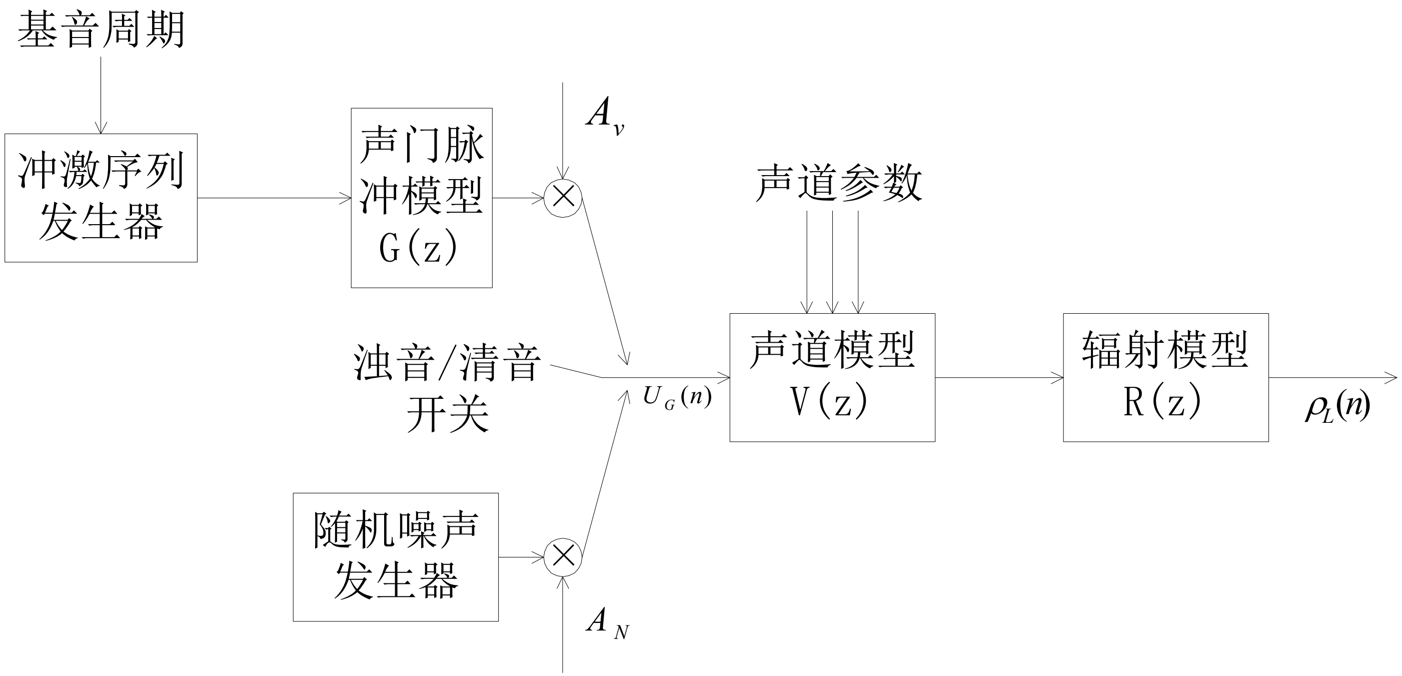


语音产生模型

语音信号产生的数字模型



语音产生模型



语音信号产生的完整模型为 $$H(z)=U(z)V(z)R(z)$$

数字语音处理III
Digital Speech Processing

https://courses.zju.edu.cn/course/join/8P84SXUIRG2

访问码: 8P84SXUIRG2

语音识别
Speech Recognition

杨莹春

yyc@zju.edu.cn

浙江大学教7-506

2021年10月8日

语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型

Applications

- Command and control
  - Manufacturing
  - Consumer products

http://www.speech.philips.com

Nuance, American Airlines:
1-800-433-7300, touch 1

- Database query
  - Resource management
  - Air travel information
  - Stock quote

- Dictation
  - http://www.lhsl.com/contacts/
  - http://www-4.ibm.com/software/speech
  - http://www.microsoft.com/speech/

Applications

Speech-recognition software

- IBM (Via Voice, Voice Server Applications,...)
  - Speaker independent, continuous command recognition
  - Large vocabulary recognition
  - Text-to-speech confirmation
  - Barge in (The ability to interrupt an audio prompt as it is playing)
- Dragon Systems, Lernout & Hauspie (L&H Voice Xpress™
- Philips
  - Dictation
  - Telephone
  - Voice Control (SpeechWave, VoCon SDK, chip-sets)
- Microsoft (Whisper, Dr Who)

Applications

The following is a transcript of a telephone conversation between Victor Zue and the Jupiter system for weather information. Jupiter's toll-free number is 1-888-573-8255.

- [Phone ringing]
- Jupiter: Hi. Welcome to Jupiter, the M.I.T. Lab for Computer Science Weather Information System. This call is being recorded for system development.
  You may hang up or ask for help at any time. How can I help you?
- V.Z.: What is the forecast for Boston today?
- Jupiter: In Boston this afternoon, partly sunny and windy. High around 50. Tonight, clear, lows around 35. What other information can I give you?
- V.Z.: Will it rain tomorrow?
- Jupiter: To my knowledge, the forecast calls for no rain in Boston tomorrow. Is there anything else?

- V.Z.: How about Seattle?
- Jupiter: To my knowledge, the forecast calls for no rain in Seattle tomorrow. Is there something else you would like to know?
- V.Z.: Is it sunny anywhere in the Caribbean?
- Jupiter: Sorry, I am not sure what you said.
- V.Z.: Is it sunny anywhere in the Caribbean?
- Jupiter: I am sorry, I may not know the city you are asking about. Try asking for the state or country. For example, what cities do you know about in Massachusetts?
- V.Z.: What is the temperature in Montego Bay?
- Jupiter: In Montego Bay today, high 86 and low 73. Is there something else?
- V.Z.: Good-bye.

语音识别的性能评价

原句：我 们 明 天 去 天 安 门

识别：我 × 明后天 去 天 坛 ×

删除错误 Deletion　　插入错误 Insertion　　替换错误 Substitution

正确率：

$$Correct = \frac{\mathrm{N - D - S}}{\mathrm{N}} \times 100\%$$

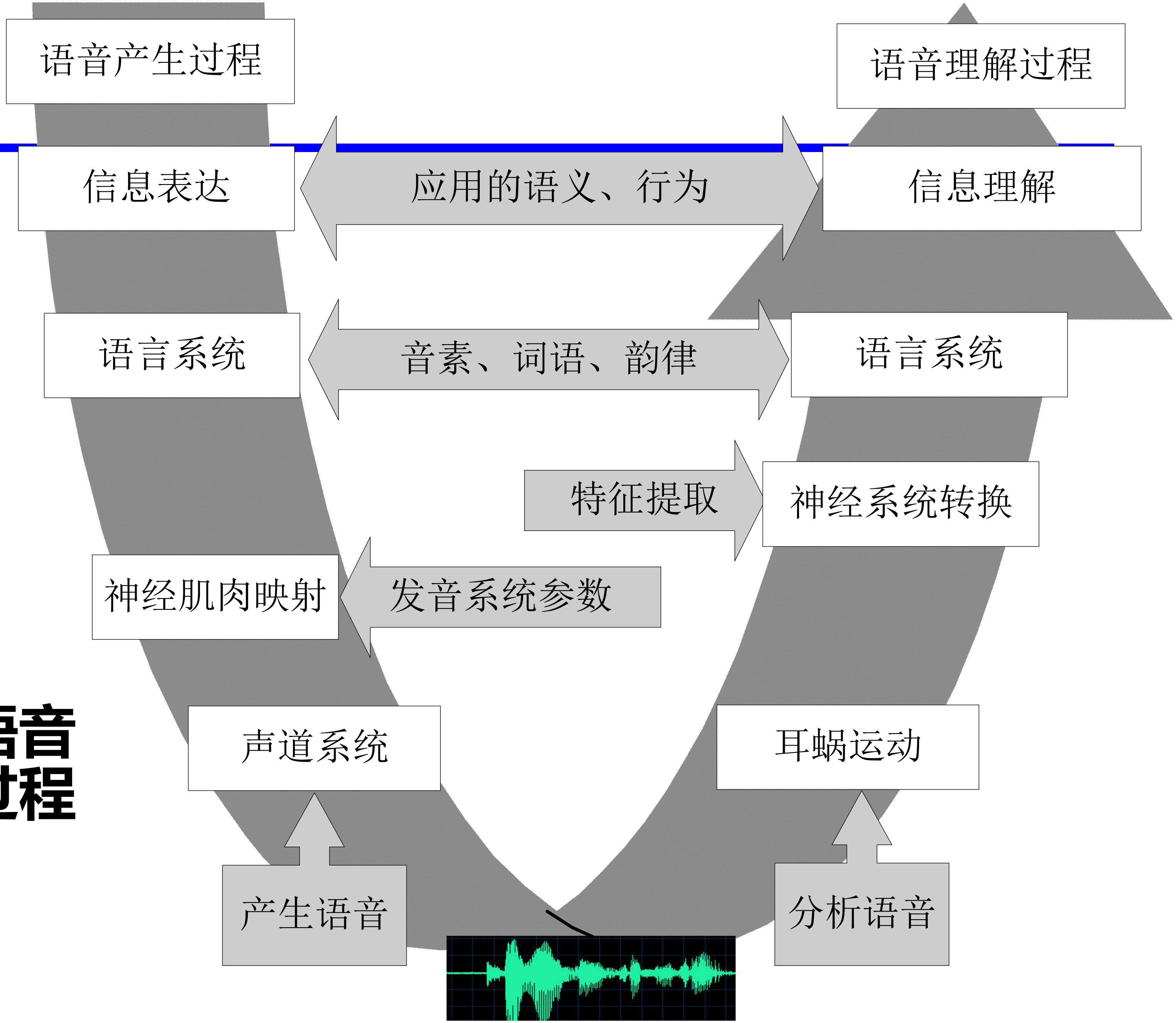
准确率：

$$Accuracy = \frac{\mathrm{N - D - S - I}}{\mathrm{N}} \times 100\%$$

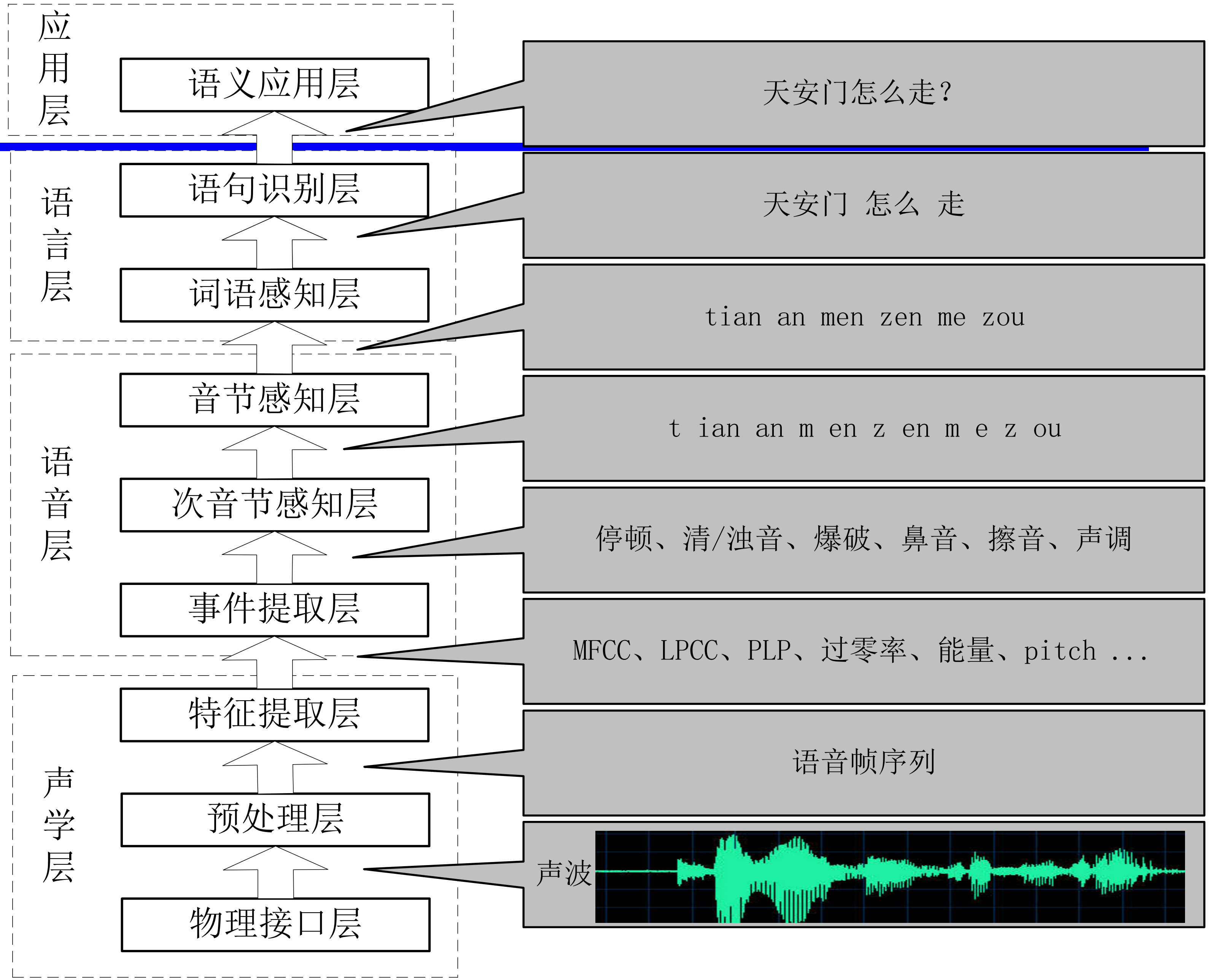
语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型

语音产生语音理解生理过程



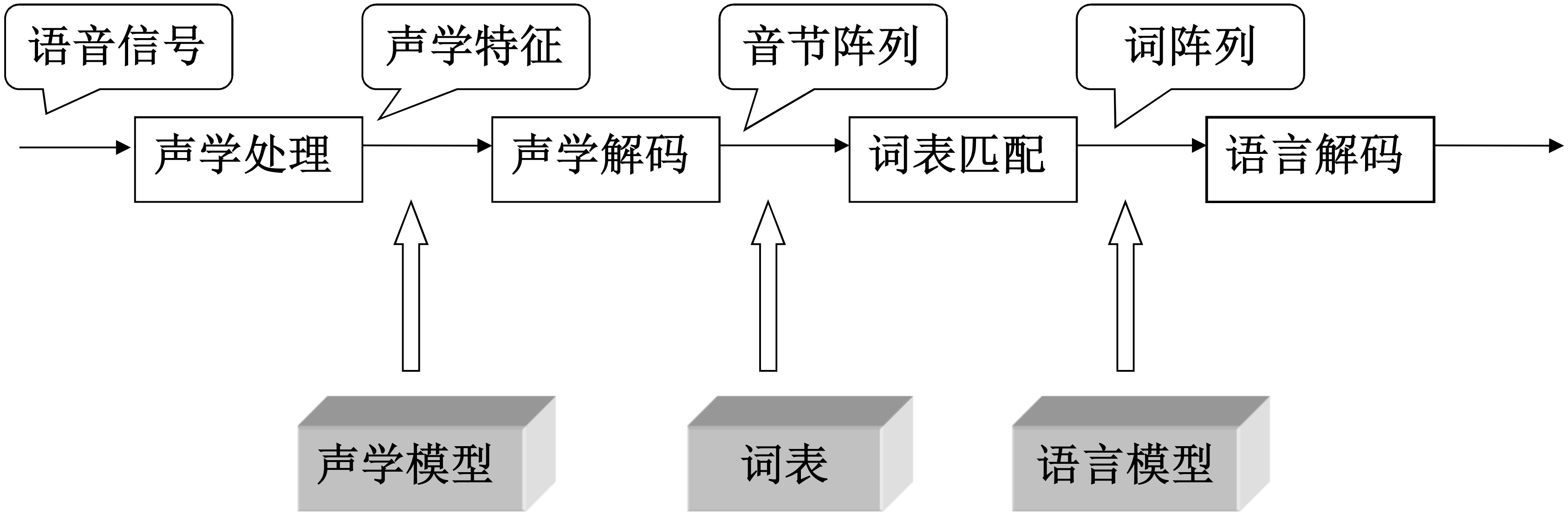
语音识别层次模型


应用层
语义应用层
天安门怎么走?
语言层
语句识别层
天安门 怎么 走
词语感知层
tian an men zen me zou
语音层
音节感知层
t ian an m en z en m e z ou
次音节感知层
停顿、清/浊音、爆破、鼻音、擦音、声调
事件提取层
MFCC、LPCC、PLP、过零率、能量、pitch ...
声学层
特征提取层
语音帧序列
预处理层
声波
物理接口层

统一层次模型——系统设计


应用层
语义应用层
将查询语句转换成查询系统中语义解释模块可接受的标准格式
语言层
语句识别层
导入语法模型，识别查询语句
词语感知层
导入该任务中的有效词汇，识别词语
语音层
音节感知层
导入单音节拼合规则，识别音节
次音节感知层
导入音素模型，识别音素
事件提取层
声学层
特征提取层
提取13维MFCC（包括C0）和一阶差分系数
预处理层
采样率16KHz、16bit线性量化、H(z)=1-0.97z-1 滤波、语音分帧（帧长20ms、帧移10ms）、Hamming窗加权、语音端点检测
物理接口层
采用强指向性抗噪麦克风、单声道录音

Statistical Speech Recognition Architectures



Turning Sounds into Words - Current Norm

X = acoustic signal sequence; W = word sequence

$P_\Lambda(W|X) = P_{\lambda_X}(X|W)\, P_{\lambda_W}(W) / P(X)$

objective: maximize the *average* performance (accuracy rate)

$\max_\Lambda P_\Lambda(W|X)$ during training

$\max_W P_\Lambda(W|X)$ during decoding

⇨

$P_{\lambda_W}(W)$

- statistical language models (mostly for large vocabulary ASR)
- grammar expressions (finite-state, context-free, ..)

$P_{\lambda_X}(X|W)$

- hidden Markov model
- mixture density - close approx. to arbitrary distribution

<u>Data-driven methods led to major advances in speech recognition.</u>

语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型

特征提取

- **预加重：** $y[n] = x[n] - \alpha \cdot x[n-1] \qquad 0.9 < \alpha < 1.0$
- **分帧：　短时平稳(10-30ms)**
- **加窗：Hamming** $w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad 0 \le n < N$
- **特征参数**
- **倒谱均值归一化**

特征参数

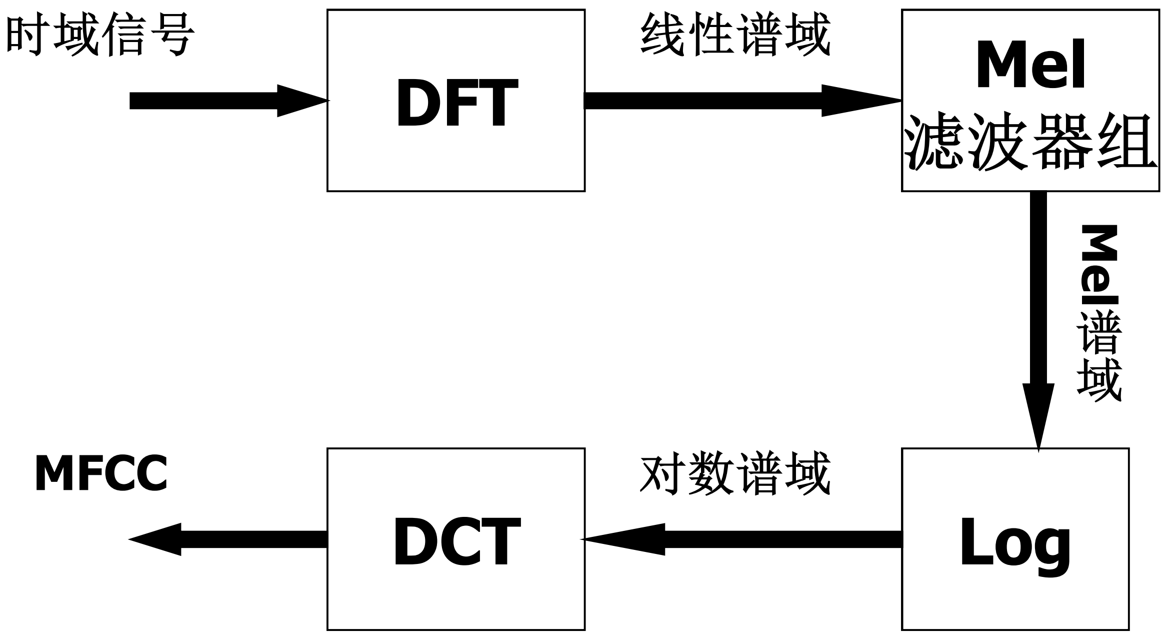
- 静态参数：Mel-Frequency Cepstrum Coefficients (MFCC)
- 帧能量
- 动态参数

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频率定义
  - 1Mel—1kHz音调感知程度的1/1000

MFCC

- **计算流程：**



Discrete Fourier Transform (DFT)

- **公式:**

$$X[k]=\sum_{n=0}^{N-1}x[n]e^{-j2\pi nk/N},0\le n<N$$

$x[n]$ -- 时域信号

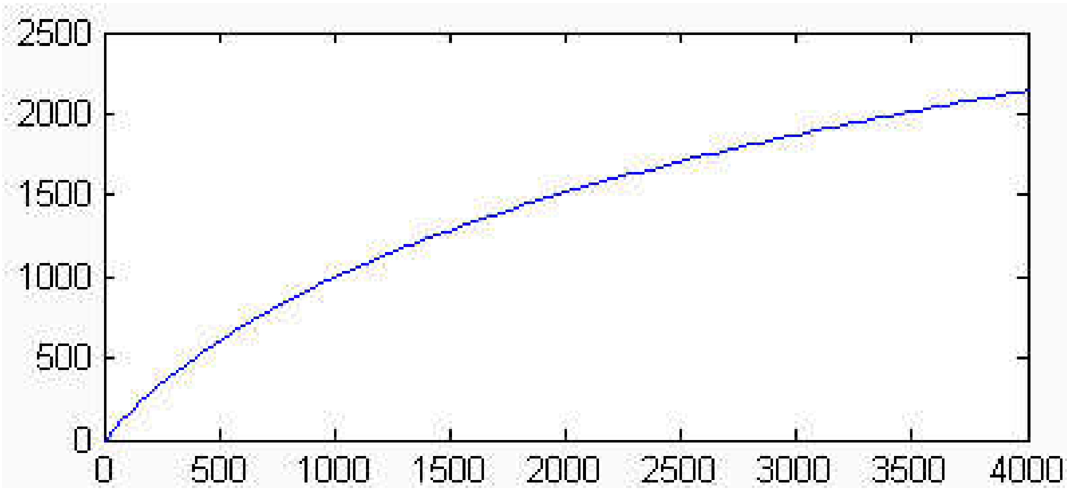
$X[k]$ -- 频域信号

Mel-频率

- **公式：**

$$B(f) = 1125\ln(1 + f/700)$$

- **频率 - Mel-频率：** f **--** 频率　　B **-- Mel-**频率



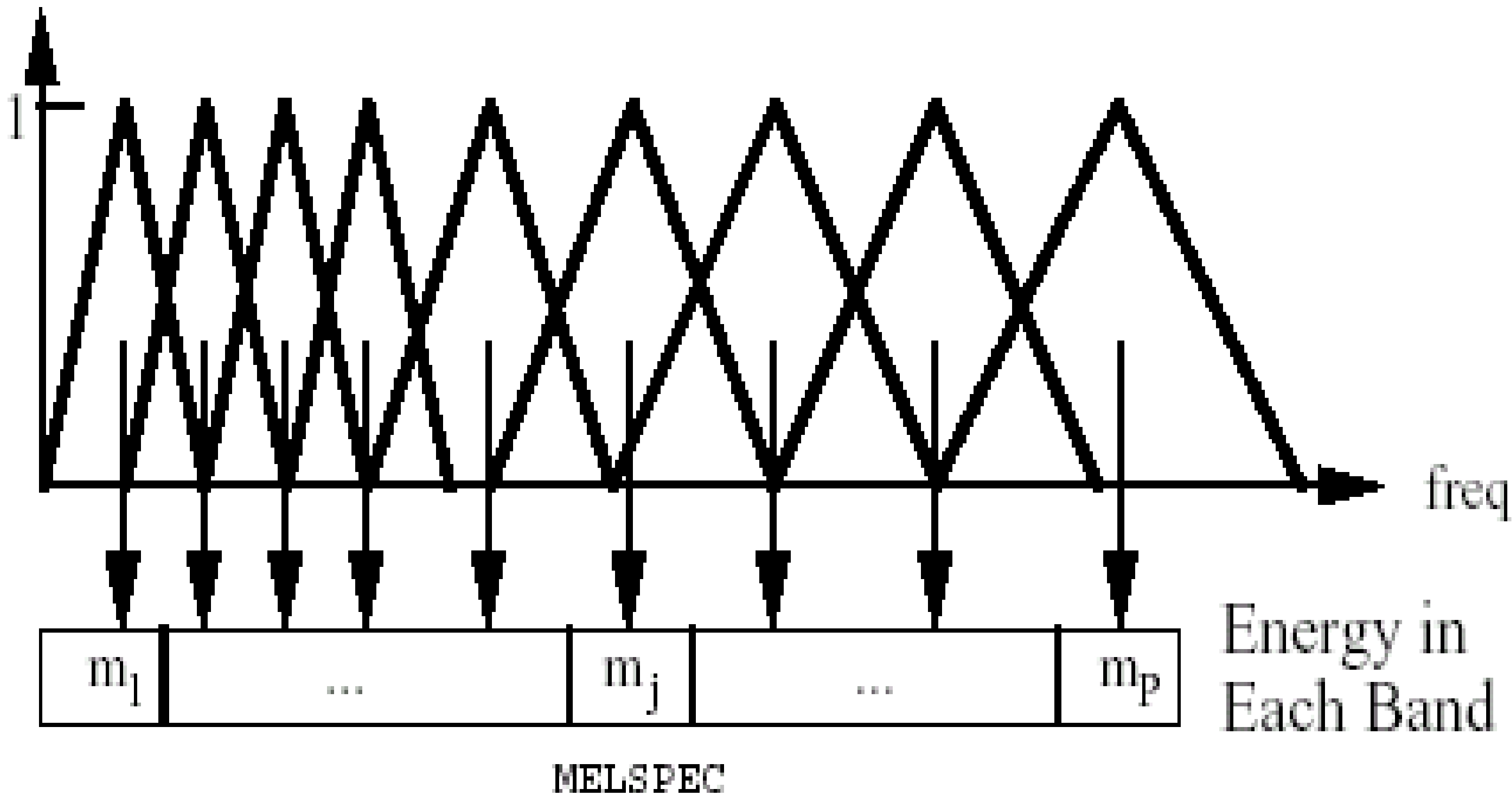
Mel-频率

频率**(Hz)**

Mel 滤波器组—参数选择

- 以采样率8kHz，帧宽30ms为例:
  - FFT窗宽：512
  - 滤波器个数：26 (通常24-40)
  - 滤波器频率应用范围（电话频带）：
    - 最高：3400Hz
    - 最低：300Hz

Mel 滤波器组—图示



对数能量

- **公式：**

$$S[m]=\ln\left(\sum_{k=0}^{N-1}|X[k]|^{2}H_{m}[k]\right)\qquad 0\le m<M$$

- **应用：对噪音和谱估计误差有更好的鲁棒性**

$$S[m]=\sum_{k=0}^{N-1}\ln\left(|X[k]|^{2}H_{m}[k]\right)\qquad 0\le m<M$$

倒谱参数

- **Discrete Cosine Transform (DCT)**

$$c[n]=\sum_{m=0}^{M-1} S[m]\cos\left(\pi n\left(m+1/2\right)/M\right) \quad 0\le n<M$$

- **倒谱维数：前12维**

帧能量

- **公式：**

$$E=\sum_{n=0}^{N-1}\left(x[n]-\overline{x}\right)^2$$

其中：$$\overline{x}=\frac{1}{N}\sum_{n=0}^{N-1}x[n]$$

- **应用：**

$$E=\sum_{n=0}^{N-1}\left|x[n]-\overline{x}\right| \qquad E=\ln\left(\sum_{n=0}^{N-1}\left(x[n]-\overline{x}\right)^2\right)$$

动态参数

- **反映帧间相关信息**
- **一阶差分：**

$$\Delta S_t = S_{t+1} - S_{t-1}$$

$$\Delta^2 S_t = \Delta S_{t+m} - \Delta S_{t-m} \qquad m = 1 或 2$$

- **二阶差分：**

S_t **--** 静态参数，包括倒谱和帧能量

倒谱均值归一化

- Cepstrum Mean Normalization (CMN)
  - 目的：消除信道带来的影响
  - 应用：T通常为整个词的特征帧数

$$\hat{O}_t = O_t - \bar{O} \qquad \text{其中} \qquad \bar{O} = \frac{1}{T}\sum_{t=1}^{T} O_t$$

- 一个变形：

$$\hat{O}_t[i] = \frac{O_t[i] - \bar{O}[i]}{\sigma[i]} \quad \text{其中} \quad \sigma[i] = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(O_t[i] - \bar{O}[i]\right)^2}$$

语音识别技术

- 发展历程
- 技术框架
- 特征提取
- 识别模型

识别模型

- 动态时间规整(DTW)
- 矢量量化(VQ)
- 隐马尔科夫模型(HMM)
- 神经网络(TDNN)
- 模糊逻辑算法

识别模型

- DTW(Dynamic Time Warping)

- VQ(Vector Quantization)

- HMM (Hidden Markov Models)

语音模型

- DTW(Dynamic Time Warping)

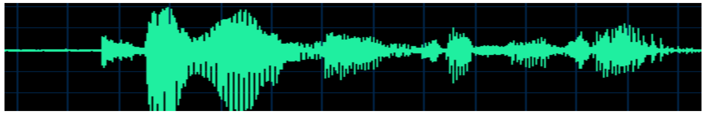
- VQ(Vector Quantization)

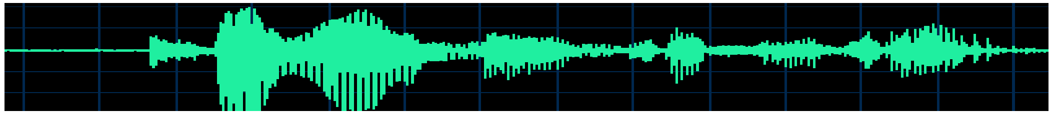
- HMM (Hidden Markov Models)

动态时间规整

- 语音识别模式匹配的问题—— 时间对准
  - 同一个人在不同时刻说同一句话、发同一个音，也不可能具有完全相同的时间长度
  - 语音的持续时间随机改变，相对时长也随机改变
- 方法1：线性时间规整
  - 均匀伸长或缩短
  - 依赖于端点检测

    通过时域分析进行，利用能量、振幅和过零率等特征
  - 缺点：仅扩展时间轴，无法精确对准

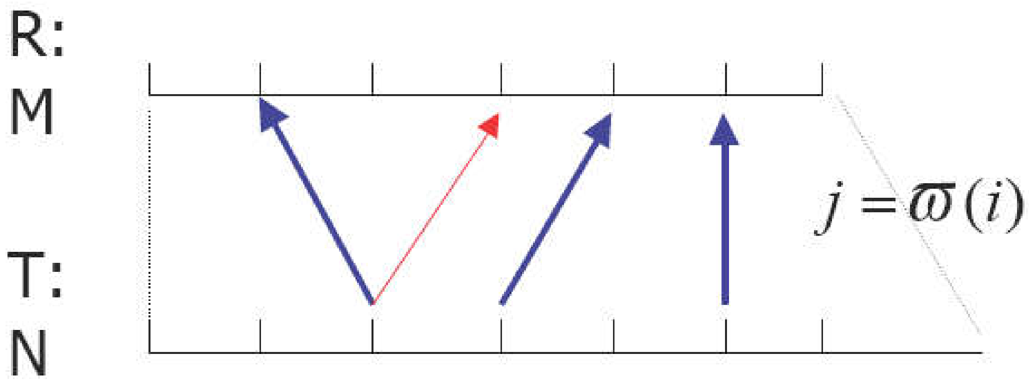




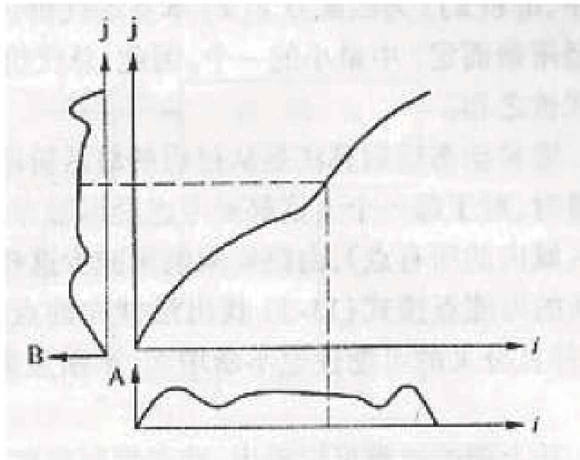
- 方法2：动态时间规整
  - DTW - Dynamic Time Warping

DTW的基本思想

- **一种非线性时间规整模式匹配算法**
  - **将时间规整与距离测度结合起来，采用优化技术，以最优匹配为目标，寻找最优的时间规整函数w(i)，从而实现大小(长短)不同的模式的比较**

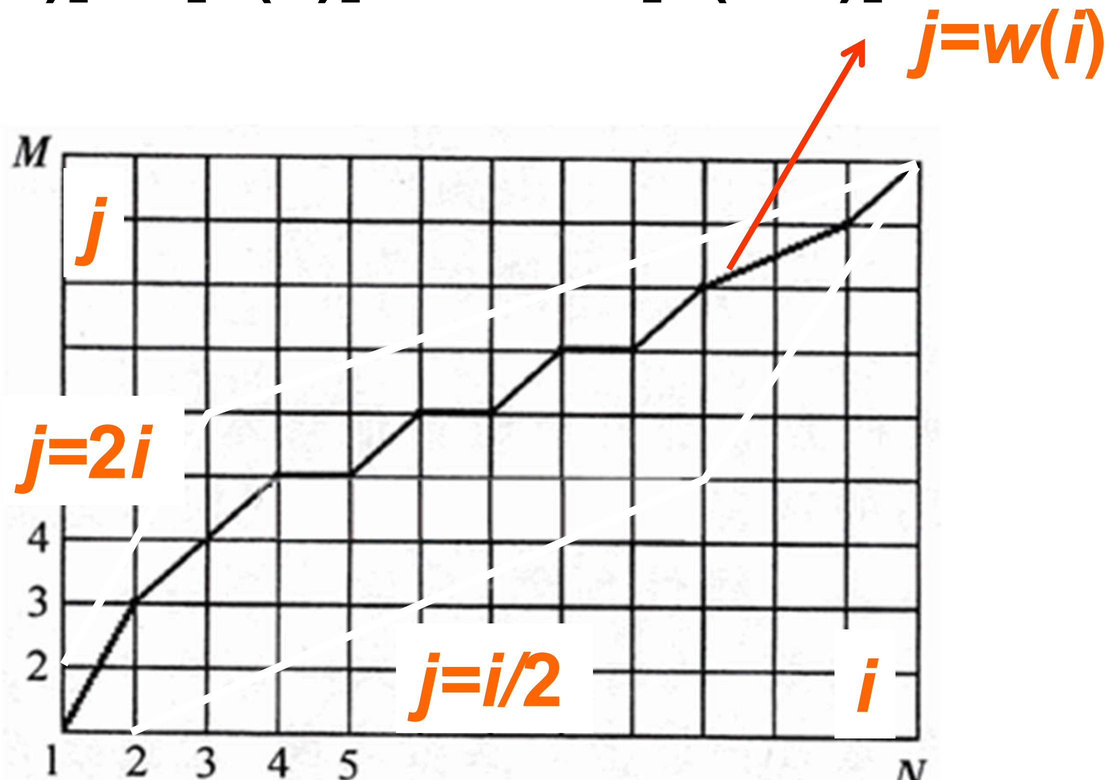


$$D=\min_{\varpi(i)}\sum_{i=1}^{M}d[T(i),R(\varpi(i))]$$



DTW的DP实现

- **动态规划**　　$D[c(k)]=d[c(k)] + \min D[c(k-1)]$
  - **搜索区域约束**
    - **平行四边形**
      - j=2i
      - j=i/2
  - **路径限制**
    - **W斜率**
      - **0，1，2**



[redacted]

DTW评价

- 适用场合
  - DTW适合于特定人、基元较少的场合
  - 多用于孤立词识别
- DTW的问题：
  - 运算量较大；
  - 识别性能过分依赖于端点检测；
  - 太依赖于说话人的原来发音；
  - 不能对样本作动态训练；
  - 没有充分利用语音信号的时序动态特性；

语音模型

- DTW(Dynamic Time Warping)

- **VQ(Vector Quantization)**

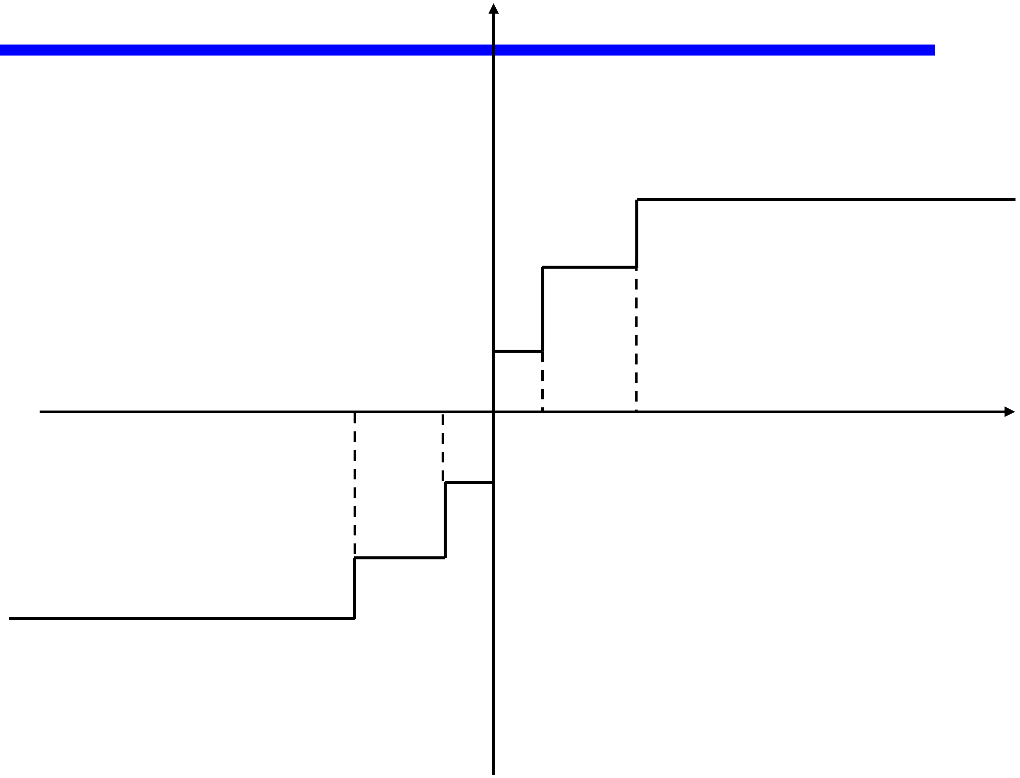
- HMM (Hidden Markov Models)

VQ在语音分析中的应用

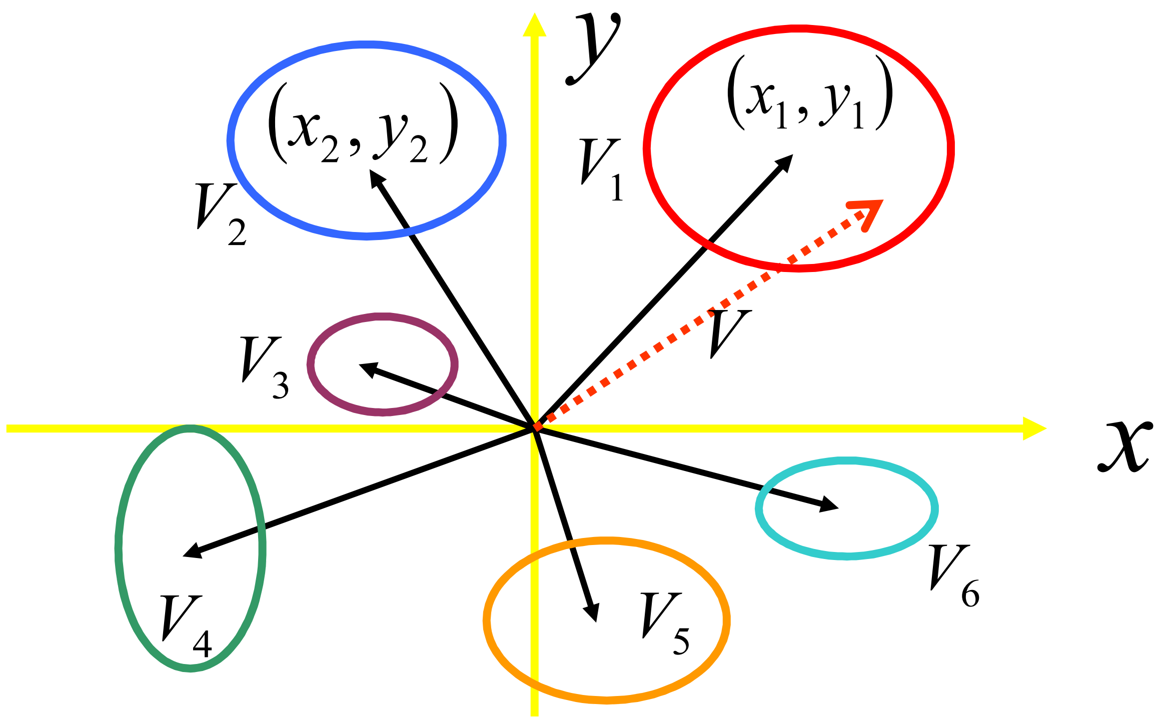
- 进入80年代以后，VQ技术引入语音处理领域，推动了语音技术发展，使之有了长足的进步
- 目前这项技术已经用于：
  - 语音识别；
  - 语音波形编码；
  - 线性预测编码；
  - ……

VQ基本概念

- 标量量化
  - 均匀
  - 非均匀
- 矢量/向量量化VQ
  - Vector Quantization
  - VQ就是将某一区域（范围）内的矢量归为某一类
- 矢量量化的基本要素
  - 聚类（Cluster）
  - 量化（Quantization）



VQ基本原理



上图的两维矢量空间里，存在6类矢量，每一类都有一个中心，称为室心(x_i, y_i)，每一室心对应一个码字矢量V_i=(x_i, y_i)，表征第i类矢量。集合{V_i}称为码本(codebook)。

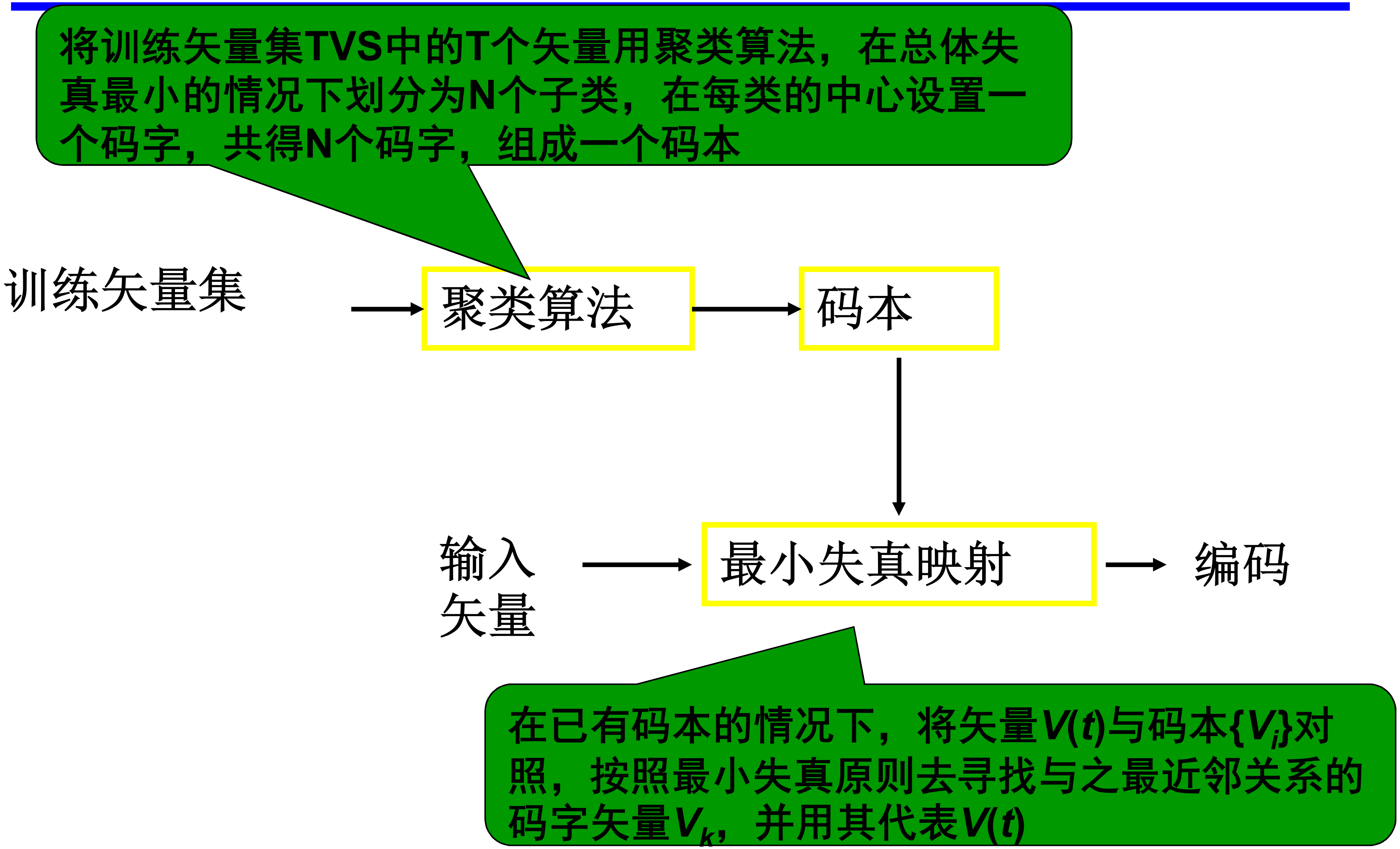
VQ基本原理

- 任意一个矢量V应该归为哪一类，要看它是“靠近”哪一类矢量，或者说它离哪一个室心最“近”
  - 例如上图中虚线画出的矢量V最靠近V1，则将其规定为V1类，并用V1表示V，或者说V被量化为V1

- 把本来无限多的矢量只用有限个码字矢量来表示
  - 上例中为6个（只需要不到3个bits表示）
  - 假如码本中的码字矢量是有序的，则被量化的矢量可用码字序号来表示。因此，可以大大压缩信息量。

VQ基本原理

- 可见VQ技术包含两个步骤
  - 先要生成码本，这是将语音的特征矢量空间首先进行划分的过程－－也称为聚类；
  - 将语音参数序列作为矢量，参照码本进行归类的过程－－也称为量化。
- 在语音处理中
  - 通常把一帧(短时窗)语音对应的特征参数（LPCC，MFCC…）用矢量表示，并称为特征矢量或特征向量；


将训练矢量集TVS中的T个矢量用聚类算法，在总体失真最小的情况下划分为N个子类，在每类的中心设置一个码字，共得N个码字，组成一个码本
训练矢量集
聚类算法
码本
输入矢量
最小失真映射
编码
在已有码本的情况下，将矢量$V(t)$与码本$\{V_i\}$对照，按照最小失真原则去寻找与之最近邻关系的码字矢量V_k，并用其代表$V(t)$

VQ的数学描述

- 假定*x*是一个*K*维向量，其各维分量都是实值随机变量。在VQ中，向量*x*要映射成另一个*K*维向量*y*，这称作把*x*量化成*y*，写作*y*=VQ(*x*)。

- *y*在一个有限集中取值，这个有限集就是一个码本，我们记作CB={CW_i: 1≤*i*≤*NC*}，*NC*为码本大小。显然，VQ的过程就是样本空间*x*到有限空间*CB*的映射：

$$x \in X \subset E^K \to y = VQ(x) \in CB \subset E^K$$

VQ的数学描述

- 当把**x**量化为**y**后，它们之间存在一个量化失真或称距离度量**d(x, y)**
- 一个量化器**VQ(·)**称为最优的是说它是所有量化器中平均/期望量化失真最小的，其中|**X**|表示集合**X**中元素的个数。

$$D = \frac{1}{|X|} \sum_{x \in X} d\left(x, VQ(x)\right)$$

VQ应用

- 在实际的实现中，某一向量**x**对某一码本**CB**量化成CW_i后，为运算方便，只用该码字在**CB**中的编号**i**来表示量化结果。这样，**VQ**可以表示为：

$$c = VQ(x) = i \text{ iff } d(x, CW_i) \le d(x, CW_j), \text{对所有} j \ne i$$

或

$$c = VQ(x) = \arg\min_i d(x, CW_i)$$

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. 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

    Ch4：4.2，4.3，4.4，4.5

    Ch5：5.1，5.6.3，5.7

    Ch9：9.1，9.2