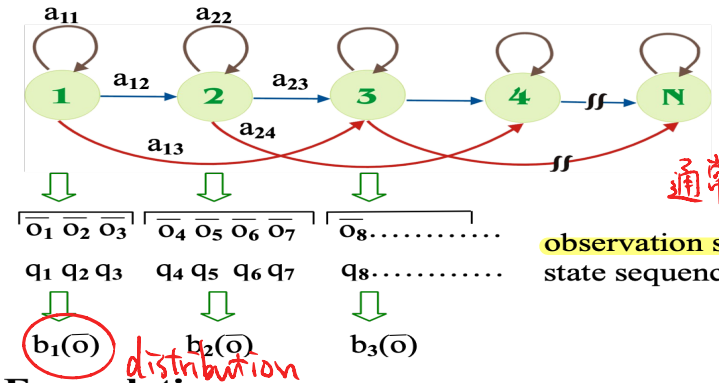
數位語音處理概論
Introduction to Digital Speech Processing

2.0 Fundamentals of Speech Recognition

References for 2.0
1.3, 3.3, 3.4, 4.2, 4.3, 6.4, 7.2, 7.3, of Bechetti

授課教師：國立臺灣大學 電機工程學系 李琳山 教授

Hidden Markov Models (HMM)



通常不reverse jump

唯一看到的東西

distribution

- **Formulation**

$\bar{o}_t = [x_1, x_2, \ldots x_D]^T$ feature vectors for a frame at time t

$q_t \in \{1,2,3 \ldots N\}$ state number for feature vector o_t

$A = [a_{ij}]$, $a_{ij} = \text{Prob}[\, q_t = j \mid q_{t-1} = i \,]$ state transition probability

多數到 $a_{i,i+2}$

e.g. a_{11} = state 1→1

$B = [b_j(\bar{o}), j = 1,2,\ldots N]$ observation (emission) probability

$b_j(\bar{o}) = \sum_{k=1}^{M} c_{jk} b_{jk}(\bar{o})$ Gaussian Mixture Model (GMM)

j-th state　k-th gaussian

$b_{jk}(\bar{o})$: multi-variate Gaussian distribution for the k-th mixture (Gaussian) of the j-th state

M : total number of mixtures

$\sum_{k=1}^{M} c_{jk} = 1$ weight of gaussion

$\pi = [\pi_1, \pi_2, \ldots \pi_N]$ initial probabilities

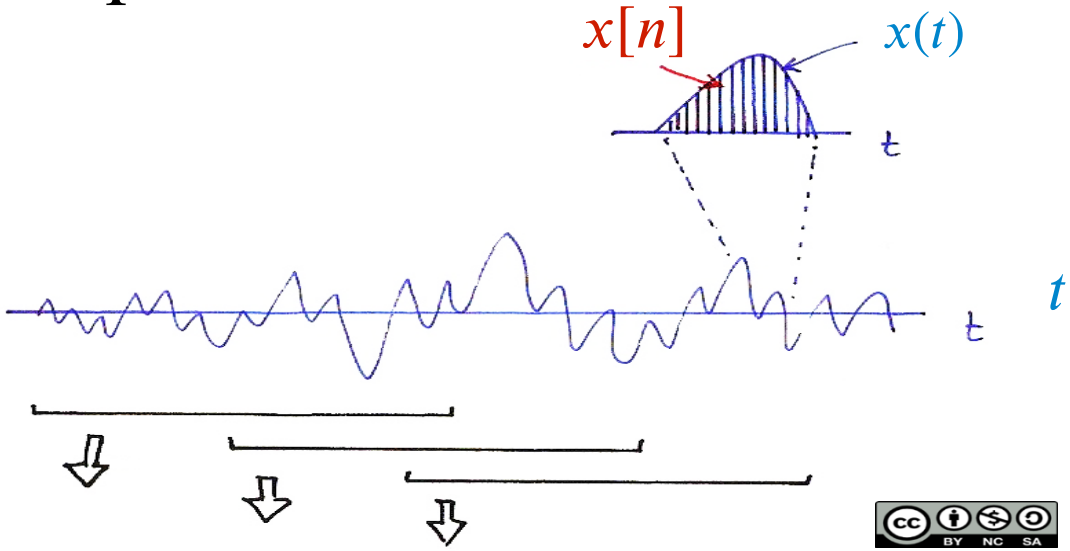
$\pi_i = \text{Prob}[q_1 = i]$ 自i開始之probability

e.g. 上 shang → set as initial

s - sh - shan - shang

$\text{HMM} : (A, B, \pi) = \lambda$

Observation Sequences



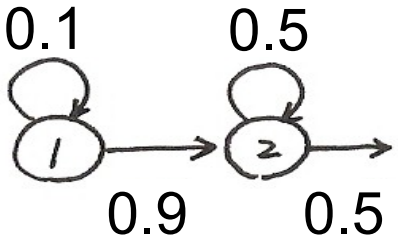
$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_D \end{bmatrix} = o_1 \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_D \end{bmatrix} = o_2 \quad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_D \end{bmatrix} = o_3$$

observation

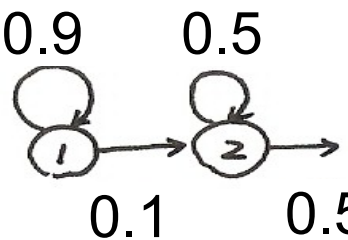
$D \simeq 39$

State Transition Probabilities

GMM: Gausion Mixture Model



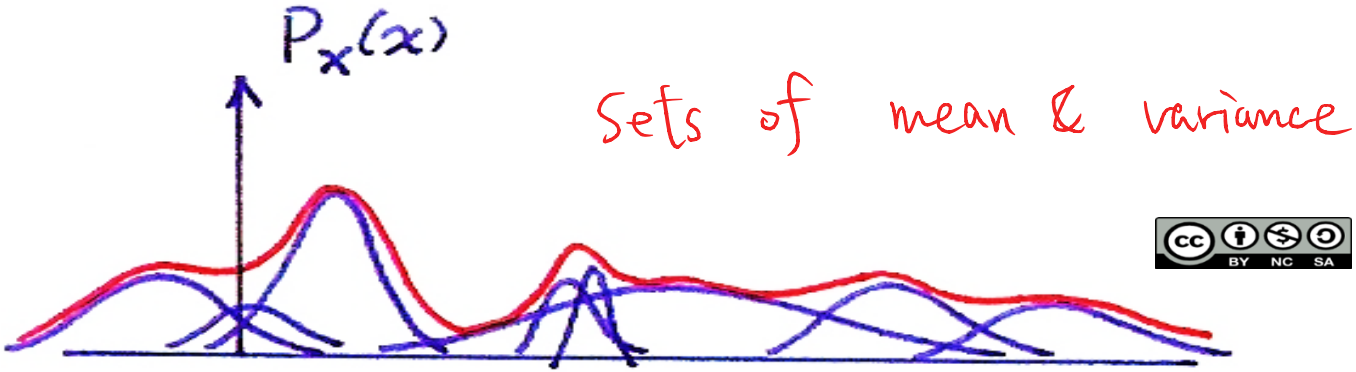
1 2 2 ⋯



1 1 1 1 1 1 1 1 2 2 ⋯

1-dim Gaussian Mixtures

probability model ⇒ use gausion to approximate



Sets of mean & variance

- **Gaussian Random Variable X**

$$f_X(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-m)^2/2\sigma^2}$$

x: value of random variable X

Random vector

- **Multivariate Gaussian Distribution for n Random Variables**

$$\overline{X} = [\, X_1, X_2, \ldots\ldots, X_n \,]^t$$

n-dimension

$$f_{\overline{X}}(\overline{x}) = \frac{1}{(2\pi)^{n/2}\Delta^{1/2}} e^{-\frac{1}{2}[(\overline{x}-\overline{\mu})^t \Sigma^{-1} (\overline{x}-\overline{\mu})]}$$

$$\overline{\mu} = [\, \mu_{X_1}, \mu_{X_2}, \ldots\ldots, \mu_{X_n} \,]^t$$

$\Sigma = [\, \sigma_{ij} \,]$, covariance matrix

$\Sigma = [\, \sigma_{ij} \,]$ (i, j)

$$\sigma_{ij} = E[\, (X_i - \mu_{X_i})(X_j - \mu_{X_j}) \,]$$

Δ : determinant of Σ

Multivariate Gaussian Distribution

$$(\vec{x}-\vec{\mu})^t \sum^{-1} (\vec{x}-\vec{\mu}) = \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}\right)^t \Sigma^{-1} (\vec{x}-\vec{\mu})$$

$$= \begin{bmatrix} x_1-\mu_1 & x_2-\mu_2 & \cdots & x_n-\mu_n \end{bmatrix} \Sigma^{-1} \begin{bmatrix} x_1-\mu_1 \\ x_2-\mu_2 \\ \vdots \\ x_n-\mu_n \end{bmatrix}$$

matrix, inverse

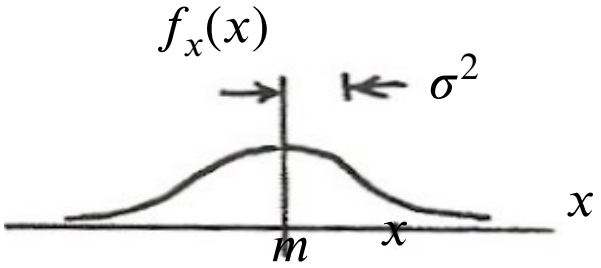
$$= (x_1-\mu_1)^2 + (x_2-\mu_2)^2 + \cdots$$

if $\Sigma = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots \end{bmatrix}$

$$= \frac{(x_1-\mu_1)^2}{\sigma_{11}^2} + \frac{(x_2-\mu_2)^2}{\sigma_{22}^2} + \cdots$$

if $\Sigma = \begin{bmatrix} \sigma_{11}^2 & & 0 \\ & \sigma_{22}^2 & \\ 0 & & \ddots \; \sigma_{nn}^2 \end{bmatrix}$

normalize ⇒ calculate the norm of probability

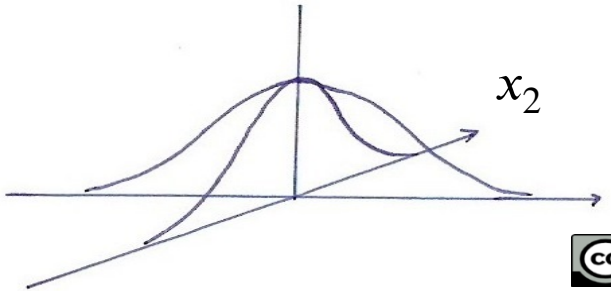


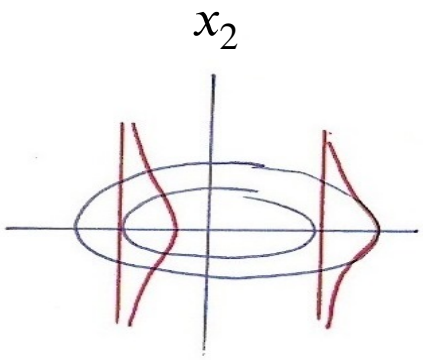
$$\sum = \begin{bmatrix} & \vdots & \\ \cdots & \sigma_{ij} & \cdots \\ & \vdots & \end{bmatrix} \begin{matrix} \\ i \\ \\ \end{matrix} \quad j$$

與 mean 差異

$$\sigma_{ij} = E[(x_i - \mu_{x_i})(x_j - \mu_{x_j})]$$

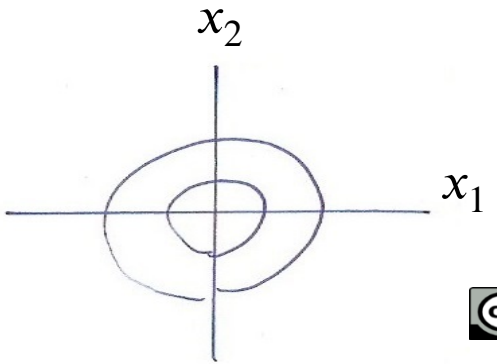
2-dim Gaussian





$$\begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2 \end{bmatrix}$$

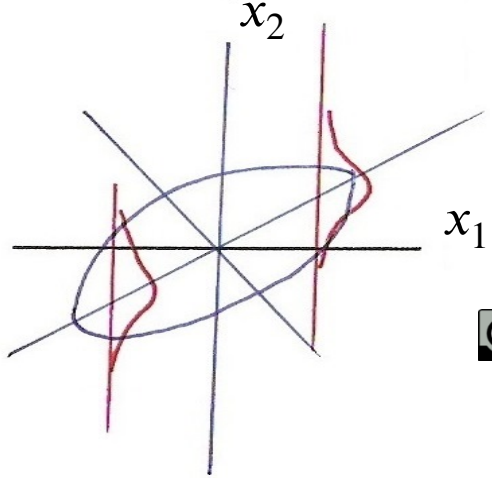
$$\sigma_{11}^2 \neq \sigma_{22}^2$$



independent

$$\begin{bmatrix} \sigma_{11}^2 & 0 \\ 0 & \sigma_{22}^2 \end{bmatrix}$$

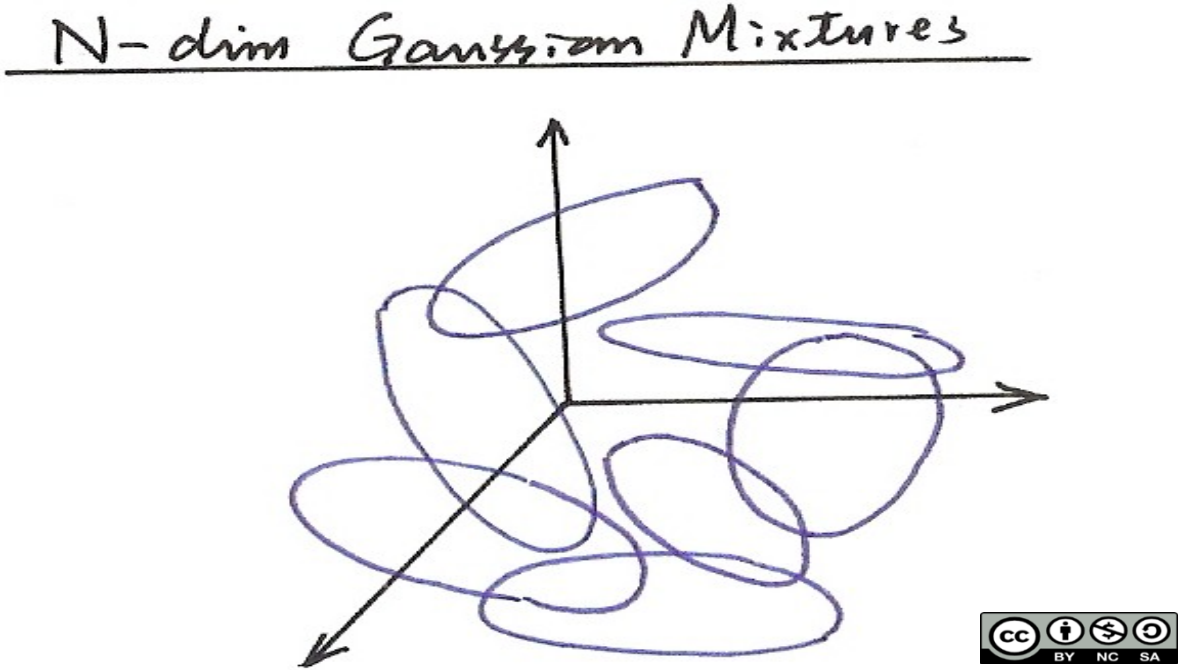
$\sigma_{11}^2 = \sigma_{22}^2$



$$\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}$$

x_1, x_2 有關

N-dim Gaussian Mixtures



Hidden Markov Models (HMM)

- **Double Layers of Stochastic Processes**
  - hidden states with random transitions for time warping
  - random output given state for random acoustic characteristics
- **Three Basic Problems**

(1) Evaluation Problem:

Given $\overline{O} = (\overline{o_1}, \overline{o_2}, \ldots \overline{o_t} \ldots \overline{o_T})$ and $\lambda = (A, B, \pi)$

find Prob [$\overline{O} \mid \lambda$]

(2) Decoding Problem:

Given $\overline{O} = (\overline{o_1}, \overline{o_2}, \ldots \overline{o_t} \ldots \overline{o_T})$ and $\lambda = (A, B, \pi)$

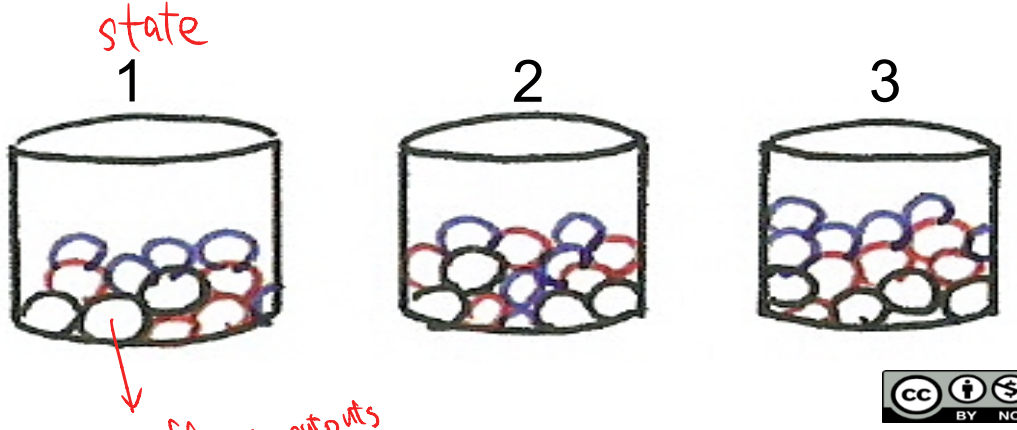
find a best state sequence $\overline{q} = (q_1, q_2, \ldots q_t, \ldots q_T)$

(3) Learning Problem:

Given $\overline{O}$, find best values for parameters in λ

such that Prob [$\overline{O} \mid \lambda$] = max

Simplified HMM



RGBGGBBGRRR……

⇒ hidden random output

from hidden random state

Feature Extraction (Front-end Signal Processing)

拉高頻率，去雜訊

- **Pre-emphasis**

  $H(z) = 1 - az^{-1}, \quad 0 << a < 1$

  $x[n] = x'[n] - ax'[n-1]$ 算對不同頻率之反應

  - pre-emphasis of spectrum at higher frequencies

- **Endpoint Detection (Speech/Silence Discrimination)**

  - short-time energy

    $$E_n = \sum_{m=-\infty}^{\infty} (x[m])^2 w[m-n]$$

    背景雜訊

    ⇒ 希望 machine 不要把 silence 作人聲

    easiest concept：算 Energy

  - adaptive thresholds

- **Windowing**

  $$Q_n = \sum_{m=-\infty}^{\infty} T\{x[m]\} w[m-n]$$

  $T\{\bullet\}$: some operator

  $w[m]$: window shape

  - Rectangular window

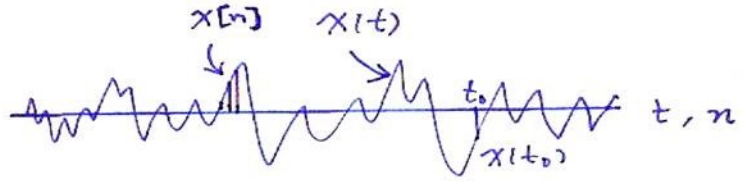
    $$w[m] = \begin{cases} 1, & 0 < m \le L-1 \\ 0, & \text{else} \end{cases}$$

    Hamming window

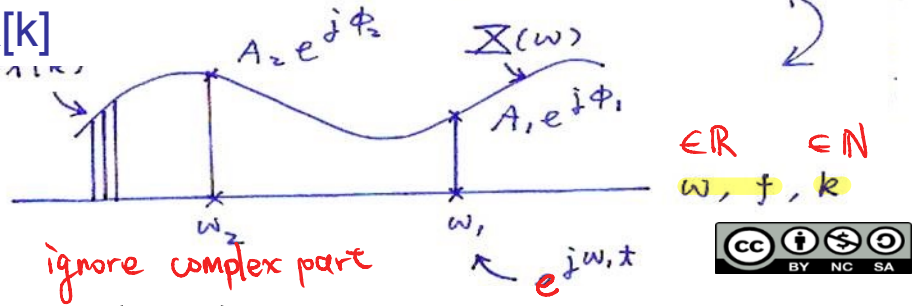
    $$w[m] = \begin{cases} 0.54 - 0.46\cos\left[\frac{2\pi m}{L}\right], & 0 \le m \le L-1 \\ 0, & \text{else} \end{cases}$$

    window length/shift/shape ← 可選擇

Time and Frequency Domains



X[k]



$$R_e\left\{e^{j\omega_1 t}\right\} = \cos(\omega_1 t)$$

$$R_e\left\{(A_1 e^{j\phi_1})(e^{j\omega_1 t})\right\} = A_1\cos(\omega_1 t + \phi_1)$$

$$\vec{X} = a_1\vec{i} + a_2\vec{j} + a_3\vec{k}$$

$$\vec{X} = \sum_i a_i x_i$$

$$x(t) = \sum_i a_i \; x_i(t)$$

把 signal 拆成一堆 vector

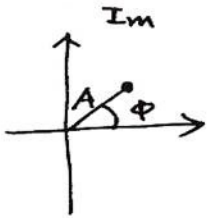
time domain

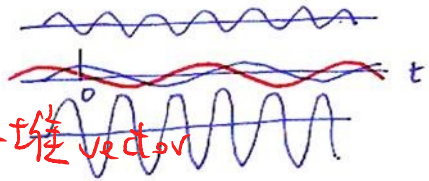
1-1 mapping

Fourier Transform

Fast Fourier Transform (FFT)

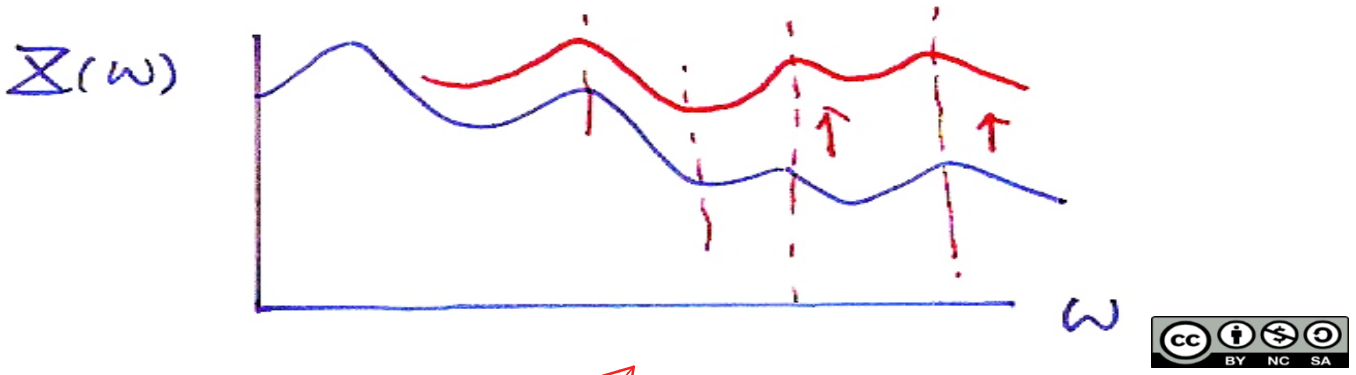
frequency domain





Pre-emphasis

∵ 很多音在高頻較弱,
∴ 才放大



- **Pre-emphasis** 放大高頻

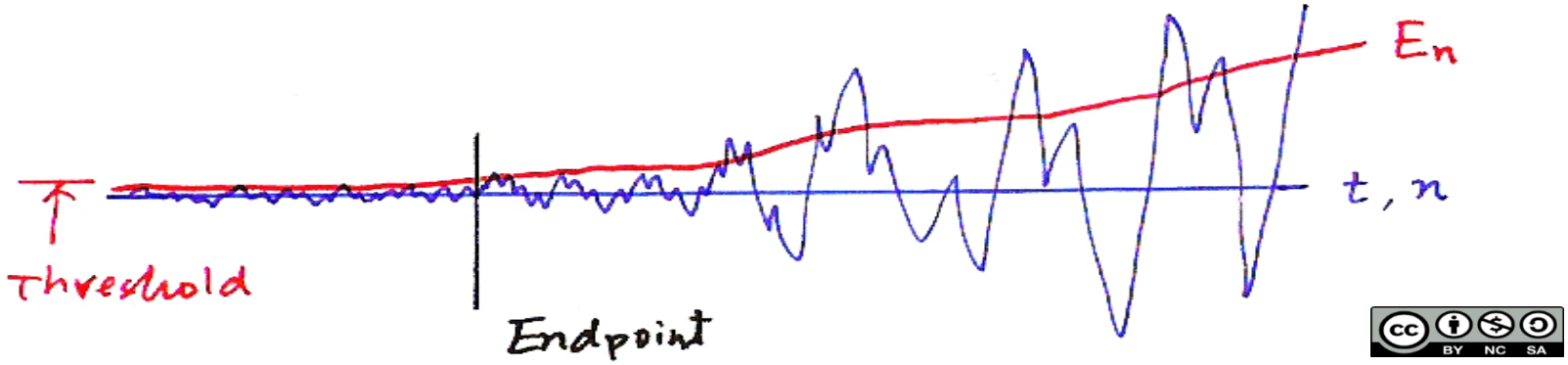
$H(z) = 1 - az^{-1}, \quad 0 << a < 1$

$x[n] = x'[n] - ax'[n-1]$

z-transform

pre-emphasis of spectrum at higher frequencies

Endpoint Detection



- **Endpoint Detection (Speech/Silence Discrimination)**
  - short-time energy

$$E_n = \sum_{m=-\infty}^{\infty} (x[m])^2 \, w[m-n]$$

window

  - adaptive thresholds

效果：hamming ≫ rectangular

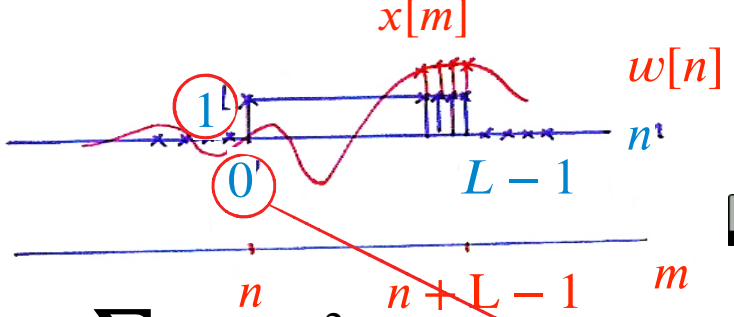
Endpoint Detection

比較smooth

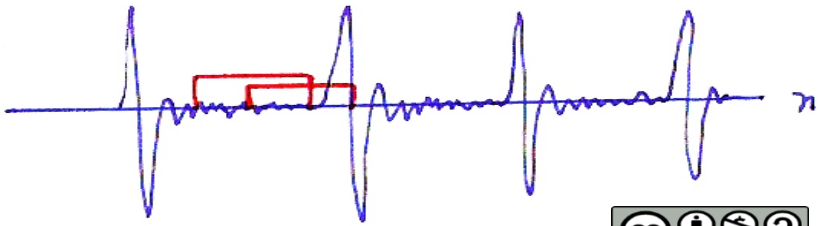
一個比較好的取window方法

⊙ Rectangular Window

設定threshold

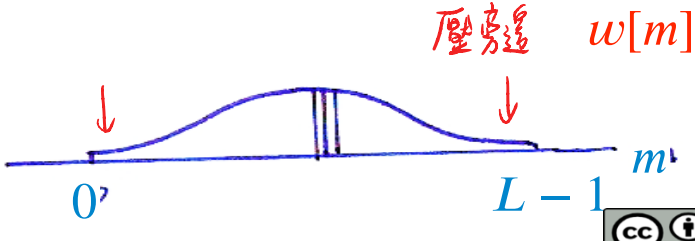


$$E_n = \sum_{-\infty}^{\infty} (x[m])^2 \, w[m-n]$$



⊙ Hamming Window

壓縮



Hamming window

$$w[m] = \begin{cases} 0.54 - 0.46\cos[\frac{2\pi m}{L}], 0 \le m \le L-1 \\ 0, \text{else} \end{cases}$$

$$Q_n = \sum_{m=-\infty}^{\infty} T\{x[m]\} \, w[m-n]$$

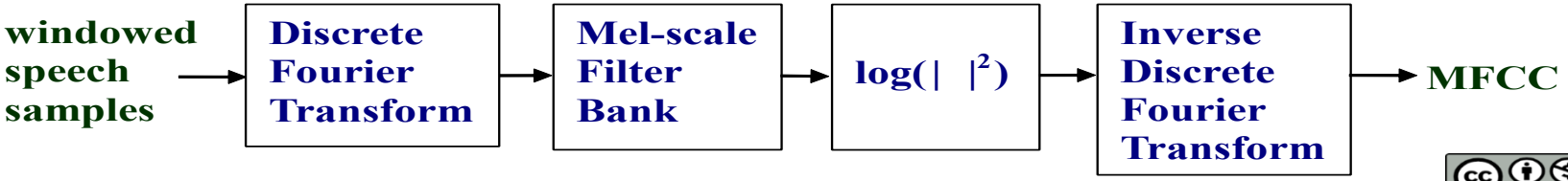
一小段一小段取，才不會波動太大

$T\{\ \bullet\ \}$: some operator

$w\{m\}$: window shape

Feature Extraction (Front-end Signal Processing)

- **Mel Frequency Cepstral Coefficients (MFCC)**



  - Mel-scale Filter Bank

    triangular shape in frequency/overlapped

    uniformly spaced below 1 kHz

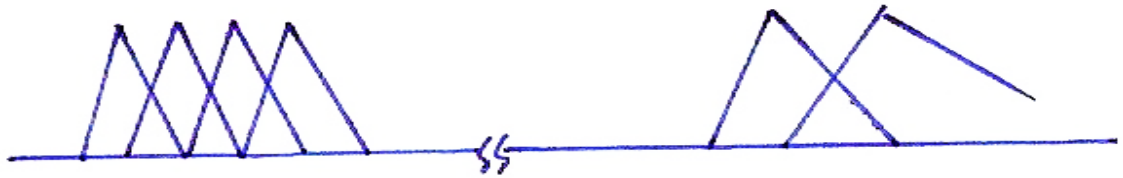
    logarithmic scale above 1 kHz 在高頻會失真且高度, 2倍 freq, 會一直變高

- **Delta Coefficients**

  - 1st/2nd order differences ∴ 高頻再取 log

14057

Mel-scale Filter Bank



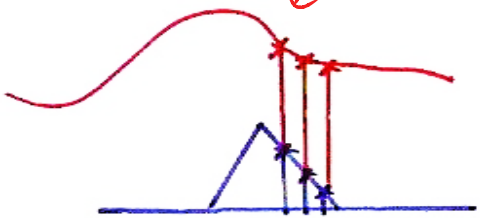
ω

low-freq

把 ω 乘上不同 weight
再加起來

$\log\omega$

high-freq



$X(\omega)$

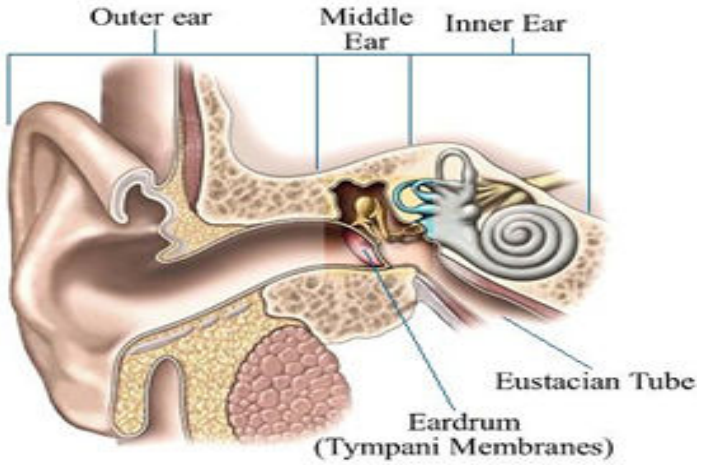
ω

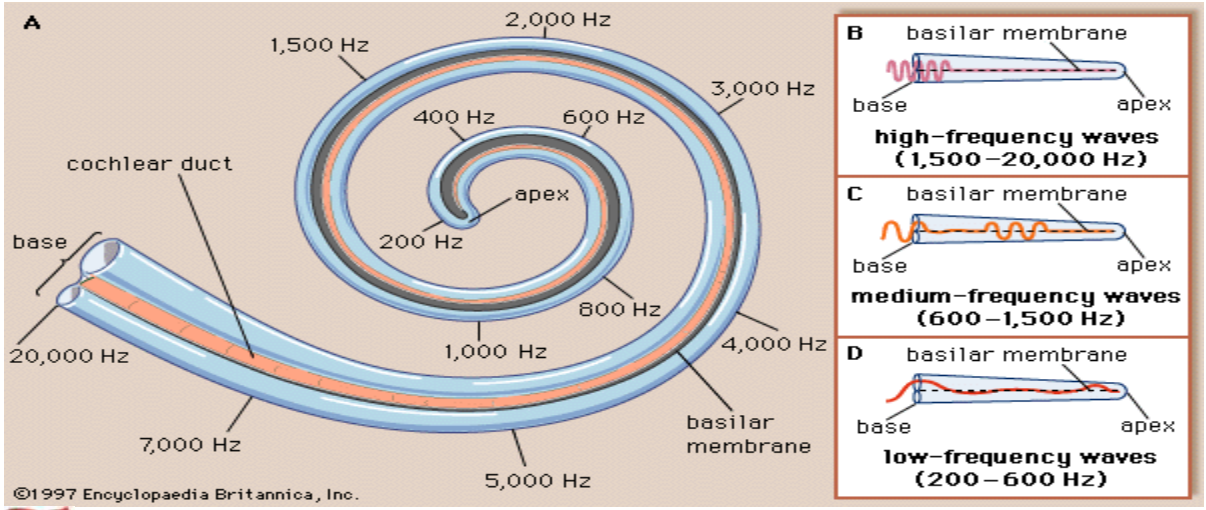
一定2個 freq
我們就不去區別

ω

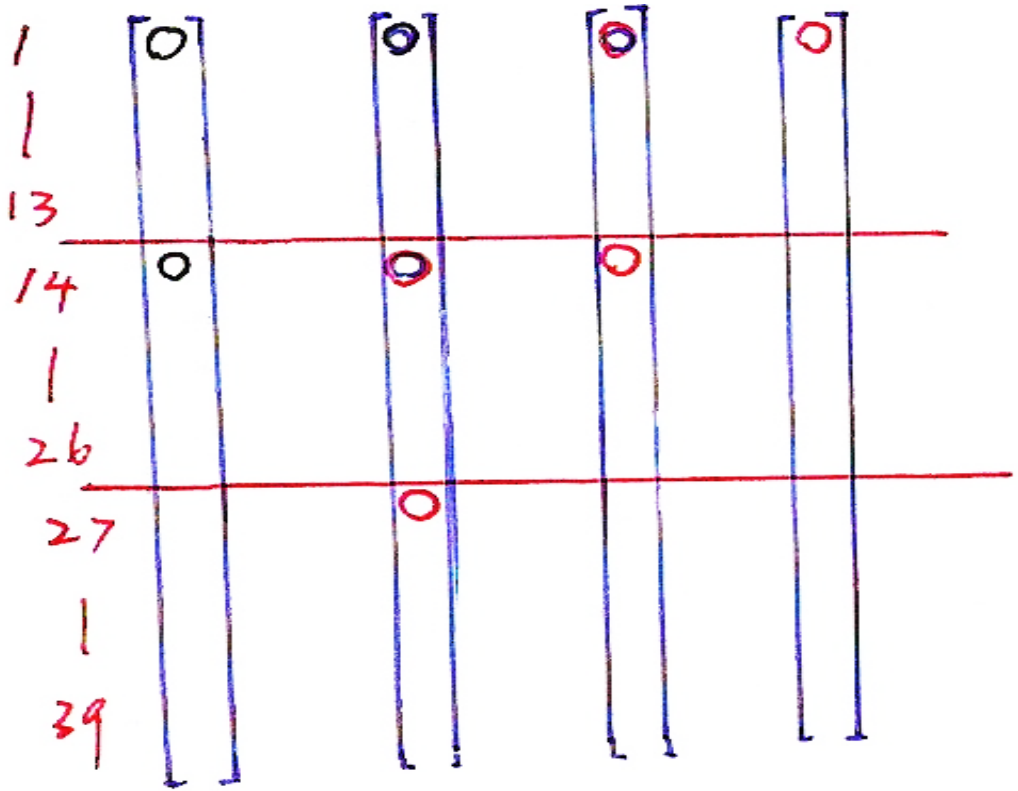
做成凸的 filter (or other shape)

Peripheral Processing for Human Perception (P.34 of 7.0)





Delta Coefficients



Language Modeling: N-gram

$W = (w_1, w_2, w_3, \ldots, w_i, \ldots w_R)$ a word sequence

- Evaluation of P(W)

$$P(W) = P(w_1) \prod_{i=2}^{R} P(w_i | w_1, w_2, \ldots w_{i-1})$$

- Assumption:

$$P(w_i | w_1, w_2, \ldots w_{i-1}) = P(w_i | w_{i-N+1}, w_{i-N+2}, \ldots w_{i-1})$$

Occurrence of a word depends on previous N−1 words only

N-gram language models

$N = 2$	:	bigram	$P(w_i \mid w_{i-1})$
$N = 3$	:	tri-gram	$P(w_i \mid w_{i-2}, w_{i-1})$
$N = 4$	:	four-gram	$P(w_i \mid w_{i-3}, w_{i-2}, w_{i-1})$
⋮			
$N = 1$	:	unigram	$P(w_i)$

probabilities estimated from a training text database

example : tri-gram model

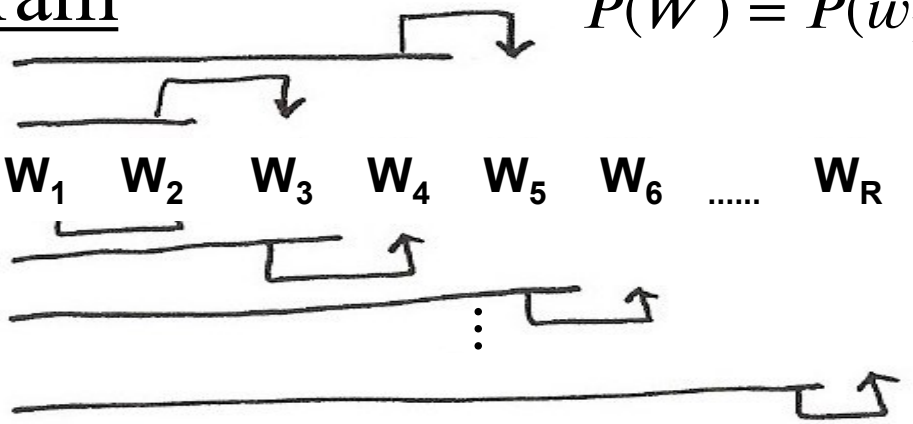
$$P(W) = P(w_1)\, P(w_2 | w_1) \prod_{i=3}^{N} P(w_i | w_{i-2}, w_{i-1})$$

N-gram

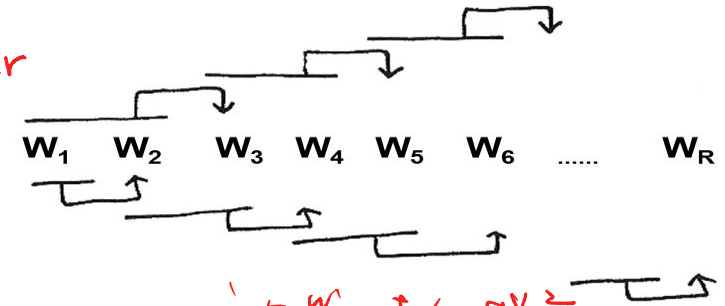
2033

$$P(W) = P(w_1)\prod_{i=2}^{R} P\left(w_i \middle| w_1, w_2, \cdots, w_{i-1}\right)$$

有1,2,…n-1個字到下個的機率



⊙ tri-grammer



$$P(W) = P(w_1)P(w_2 \mid w_1)\prod_{i=3}^{N} P\left(w_i \middle| w_{i-2}, w_{i-1}\right)$$

前2個后到下的prob

只算3連的機率

∵太複雜，但貧窮不通，只是方便有效

Language Modeling

- Evaluation of N-gram model parameters

unigram

$$P(w^i) = \frac{N(w^i)}{\sum_{j=1}^{V} N(w^j)}$$

w^i : a word in the vocabulary

V : total number of different words in the vocabulary

N(·) number of counts in the training text database

bigram

$$P(w^j|w^k) = \frac{N(<w^k,w^j>)}{N(w^k)}$$

$< w^k, w^j >$: a word pair

trigram

$$P(w^j|w^k,w^m) = \frac{N(<w^k,w^m,w^j>)}{N(<w^k,w^m>)}$$

smoothing – estimation of probabilities of rare events by statistical approaches

... this	50000
... this is	500
... this is a ...	5

Prob [is| this] = $\frac{500}{50000}$

Prob [a| this is] = $\frac{5}{500}$

bigram

$$P(w^j | w^k) = \frac{N(\langle w^k, w^j \rangle)}{N(w^k)}$$

$\langle w^k, w^j \rangle$: a word pair

trigram

$$P(w^j | w^k, w^m) = \frac{N(\langle w^k, w^m, w^j \rangle)}{N(\langle w^k, w^m \rangle)}$$

Large Vocabulary Continuous Speech Recognition

$W = (w_1, w_2, \ldots w_R)$ a word sequence

$\overline{X} = (\overline{x}_1, \overline{x}_2, \ldots \overline{x}_T)$ feature vectors for a speech utterance

$$W^* = \underset{W}{\text{Arg Max}} \ \text{Prob}(W|\overline{X})$$ MAP principle

given ō, find W̄'s prob

$$\text{Prob}(W|\overline{X}) = \frac{\text{Prob}(\overline{X}|W) \cdot P(W)}{P(\overline{X})} = \max$$

ignore, ∵ same

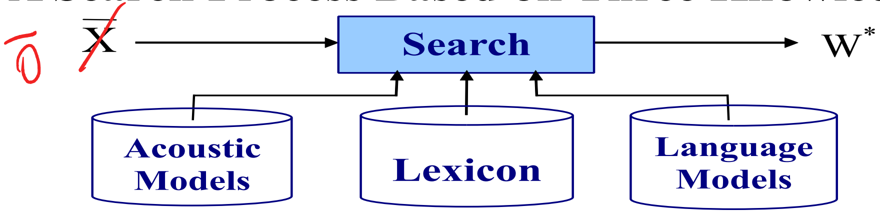
A Posteriori Probability (= W*)

⇒ Maximum A Posteriori (MAP) Principle

$$\text{Prob}(\overline{X}|W) \cdot P(W) = \max$$

by HMM — by language model

- A Search Process Based on Three Knowledge Sources



把小model串成大model

- Acoustic Models : HMMs for basic voice units (e.g. phonemes)
- Lexicon : a database of all possible words in the vocabulary, each word including its pronunciation in terms of component basic voice units
- Language Models : based on words in the lexicon

Maximum A Posteriori Principle (MAP)

$W : \{ w_1 , w_2 , w_3 \}$

↑ ↑ ↑

sunny rainy cloudy

$P(w_1)$
$P(w_2)$
$+ P(w_3)$

1.0

$\vec{X} = (x_1, x_2, x_3 \ 3 \ldots\ldots)$ weather parameters

⊙ Problem

given $\vec{X}$ today, to predict W for tomorrow

Maximum A Posteriori Principle (MAP)

⊙ Approach 1

Comparing $P(w_1)$, $P(w_2)$, $P(w_3)$

$\vec{X}$ not used?

⊙ Approach 2

A Posteriori Probability 事後機率

Likelihood function

Prior Probability 事前機率

compute $i = 1, 2, 3$

$$P(w_i \mid \vec{x}) = \frac{P(\vec{x} \mid w_i) \cdot P(w_i)}{P(\vec{x})}, \quad P(\vec{x}) = \sum_{w_1}^{w_3} P(\vec{x} \mid w_i) \cdot P(w_i),$$

unknown → observation

compare $P(w_2 \mid \vec{x})$ with $P(w_3 \mid \vec{x})$

likelihood function 現象的 distribution　　prior probability

$$P\left(\begin{matrix}w_1\\w_2\\w_3\end{matrix}\middle|\bar{x}\right)=\frac{P\left(\bar{x}\middle|\begin{matrix}w_1\\w_2\\w_3\end{matrix}\right)\cdot P\left(\begin{matrix}w_1\\w_2\\w_3\end{matrix}\right)}{P(\bar{x})}$$

$$P(w_i|\bar{x})=\frac{P(\bar{x}|w_i)P(w_i)}{P(\bar{x})}$$

$i=1\sim3$　　doesn't matter

unknown　　observation　　現在所有 parameter

目的：事後機率 A posteriori probability　⇒ Compare $P(\bar{x}|w_i)\cdot P(w_i)$

2:02:00 : CH3 Map

Syllable-based One-pass Search

- **Finding the Optimal Sentence from an Unknown Utterance Using 3 Knowledge Sources:Acoustic Models, Lexicon and Language Model**
- **Based on a Lattice of Syllable Candidates**