installation Tensorflow

CPU GPU TPU

librispeech SER - Sentence error rate

WER - word error rate

WSJ - wall street journal

corpus - collection of trained dataset

librivox - open source platform where get speech dataset

openSLR -

Daniel Povey

100, 360, 500

LM language Model

Acoustic Model (it is trained Model WSJ)

related to sound

Subset	hours	per sp min	Male	femal	total
train.clean.100	100.6	25	126	125	251

Automatic Speech ~~Rego~~ Recognition

DTW . GMM, HMM DNN

Processing Audio Signal

extraction

Techniques

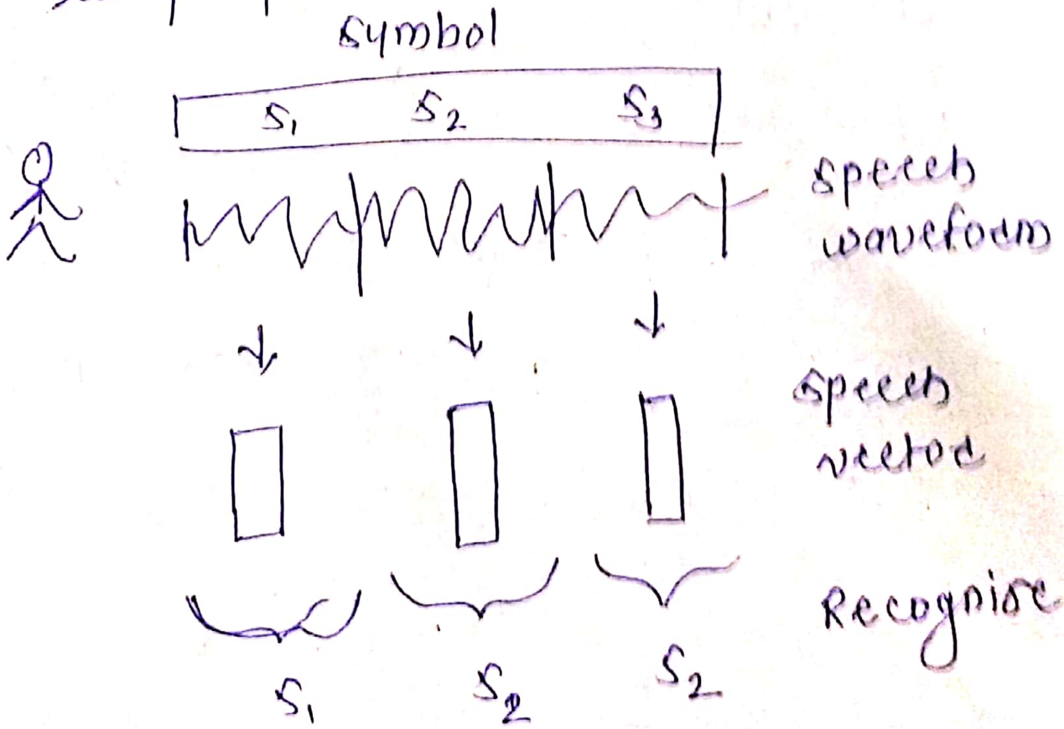
Statistical Model - GMM

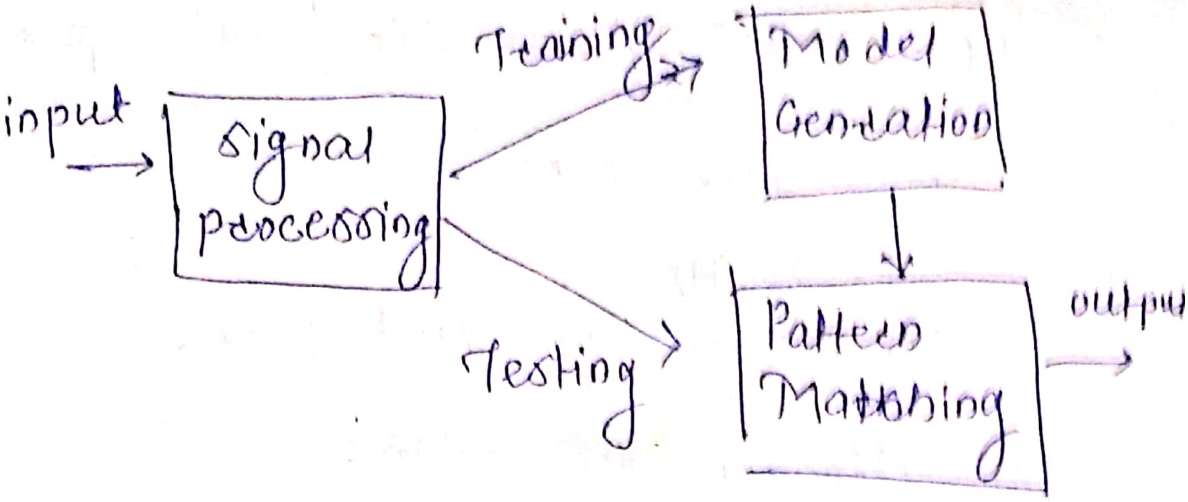
Recognition word - DTW

Recognition sentence - HMM

DNN - HMM

language Model for SR





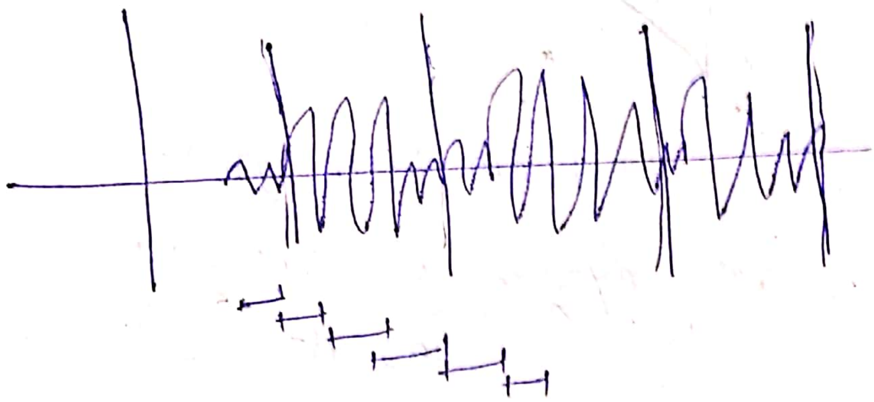
Training - learning
Testing - Recognise

* Short time processing of SR

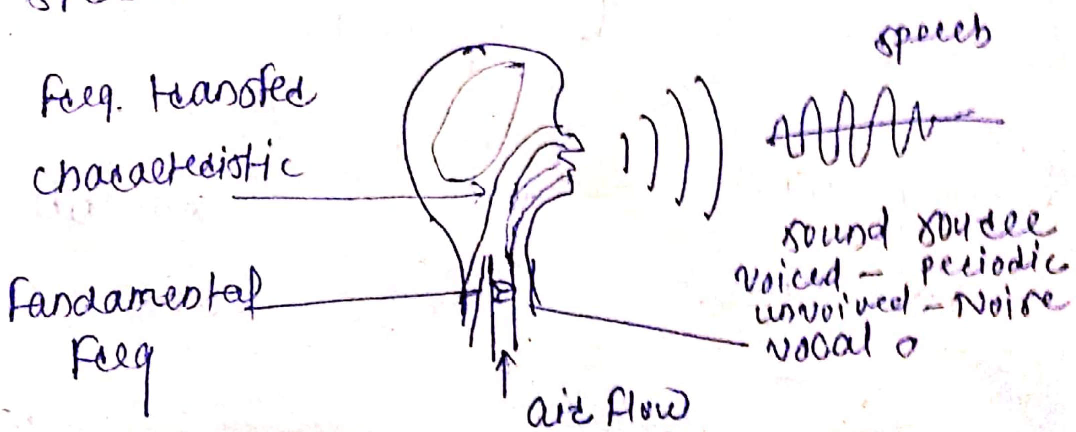
perform frequency analysis of short segment

Frame size

Frame shift



Speech Production Mechanism



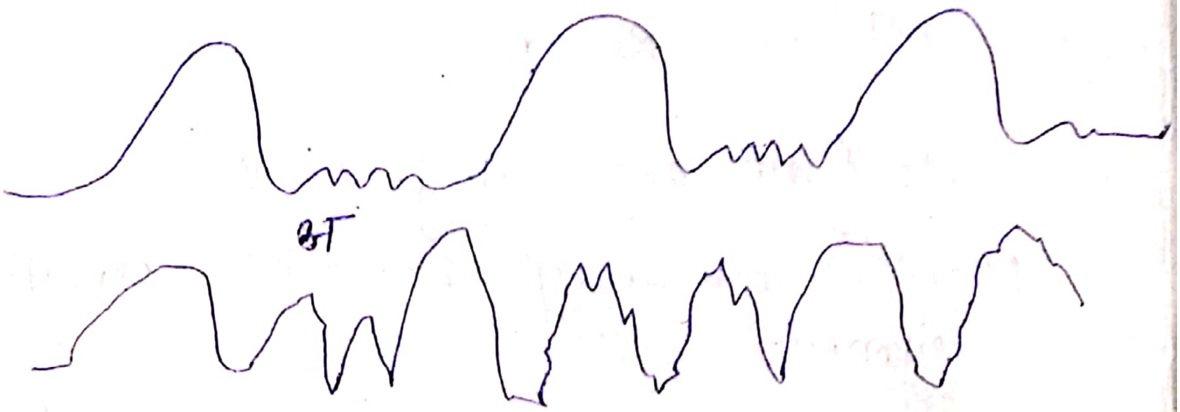
input sound is periodic / seq. of pulse

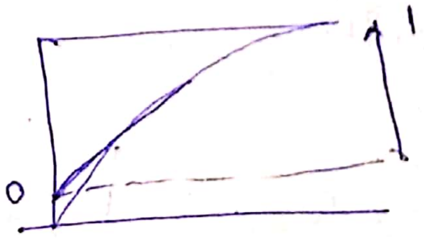
pitch frequency

300 Hz

Female is high pitch freq.

Production of voiced sound





uniform tube model

$$v = c/\lambda = 34000/4 * 17$$

$$= 500\ Hz$$

source → Filter → output

glottal vibration — vocal tract — speech

$$s(n) = e(n) * h(n)$$

exitation signal

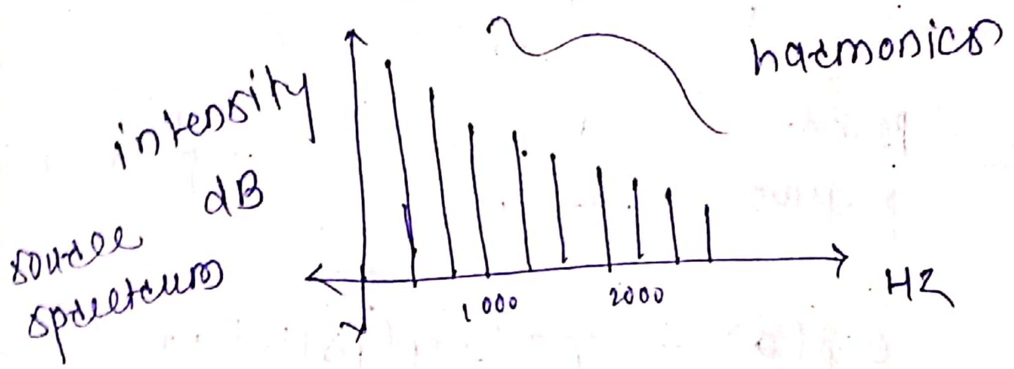
convolution

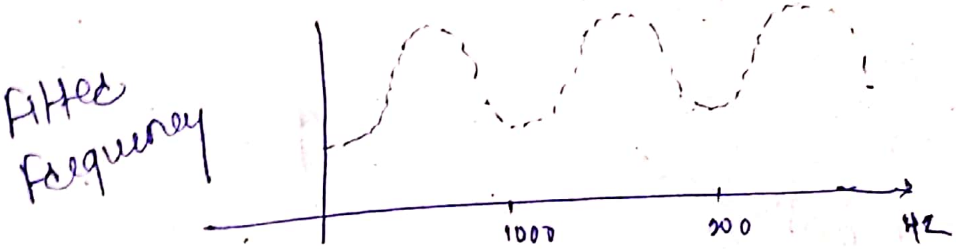
impulse

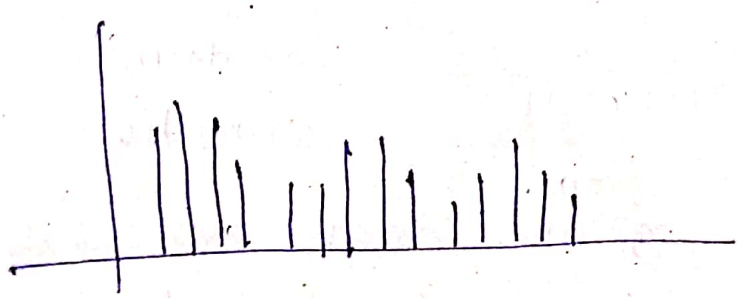
$$S(k) = E(k)\,H(k)$$

$$\log(|S(k)|**2) = \log(|E(k)**2) + \log(|H(k)|**2)$$

exitation signal is periodic





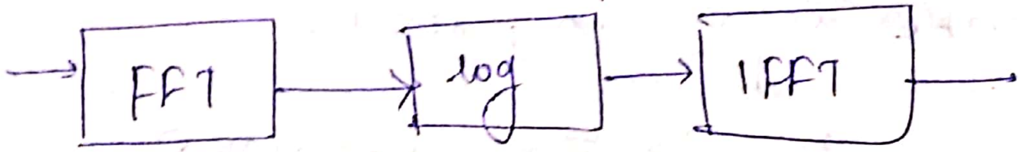


output energy spectrum

Glottal air flow

vocal tract

time waveform



wave form

power spectrum

log spectrum.

~~cepte~~ cepstrum

$$cep(q) = IFFT(\log(|S(k)|**2)$$

$q = 0, 1, \ldots N-1$

formation

peaks in spectrum

कॉक्लिया

ear drum vibration

3 cm
30 mm

20 Hz - 20,000 Hz

audio freq.

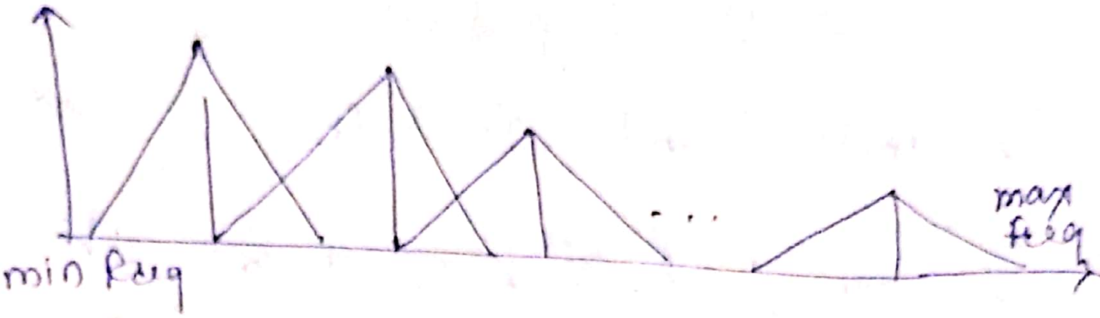
Basilar membrane

Bark scale / mel scale

Half of part is linear

& the half is logarithmic

MFCC



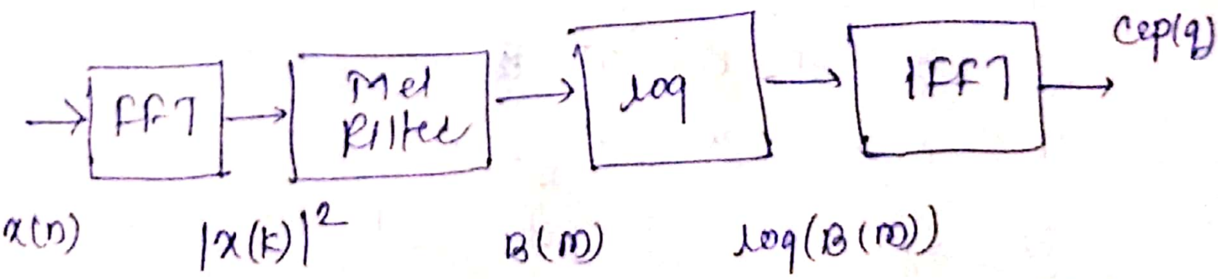
triangles = # mel filters = length of mel spectrum

$$B(m) = \sum_{k=lo(m)}^{hi(m)} |X(k)|^2$$

$$cep(q) = IFFT\{\log(|B(m)|^2)\}$$

$q = 0, 1, \ldots N$

mel frequency cepstral coefficients



extraction of features (MFCCs) that will be used for representation as well as recognition of speech sound.

Acoustic phonetics

phones and phonemes

human language

smallest meaningful contrastive unit

allphones = p & ph

Aspirate Sound प & फ

अ	आ	इ	ई	उ	ऊ	ए	ऐ	ओ	औ
a	A	i	I	u	U	e	E	o	O

क	ख	ग	घ	ङ
k	kh	g	gh	ng
च	छ	ज	झ	ञ
c	ch	j	jh	nj
ट	ठ	ड	ढ	ण
T	Th	D	Dh	N
त	थ	द	ध	न
t	th	d	dh	n
प	फ	ब	भ	म
p	ph	b	bh	m

य	र	ल	व	श	ष
y	r	l	w	sh	
स	ष	ह	क्ष	त्र	ज्ञ
s	S	h			

delta coefficient

$$y(x) = mx + c$$

$$\Delta cep(n,l) = \frac{\sum_{l=-L}^{L} l\, cep(n,l)}{\sum_{l=-L}^{L} (l)^2}$$

sequence of feature vector

Digitisation of analog speech signal

Blocking signal into frames

~~39~~ 13 FFT → mel filter → log → IFFT → MFCC

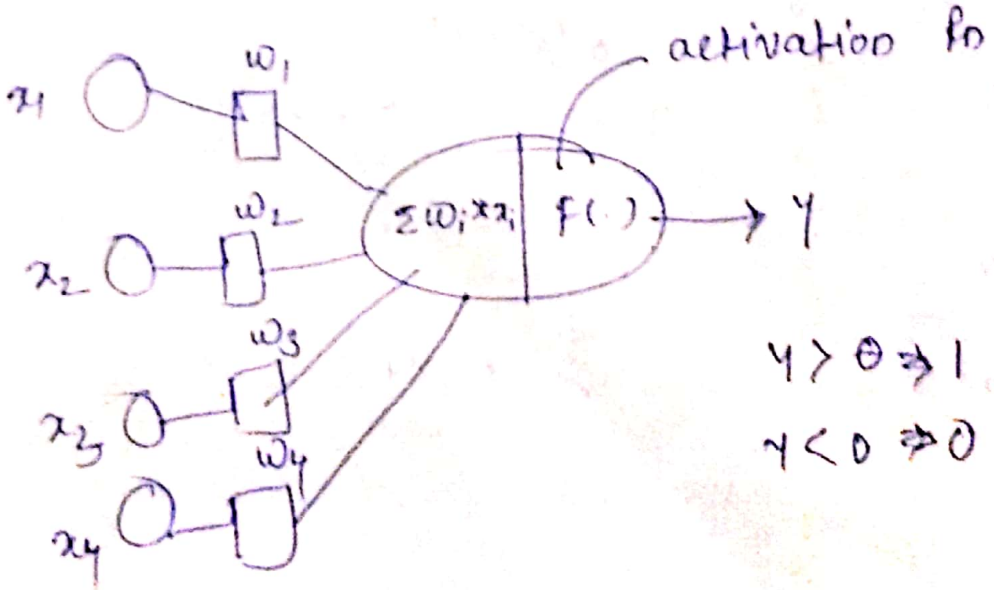
13+13 Slope and ~~vee~~ curvature

sequence of feature vectors

: $x_1, x_2 \dots x_T$ 39-dim

: $O_1, O_2 \dots O_T$ $\boxed{T = 150}$

linear perception



$y > \theta \Rightarrow 1$

$y < 0 \Rightarrow 0$

loss / cost function

$$f(w) = 0.5 * (t(n) - y(n))^2$$

where,

$$y(n) = \text{Sum}(w_i * x_i)$$

$$w_{new} = w_{old} - \frac{df}{dw}$$

A/c ~~to~~ gradient decent algorithm, weight update rule,

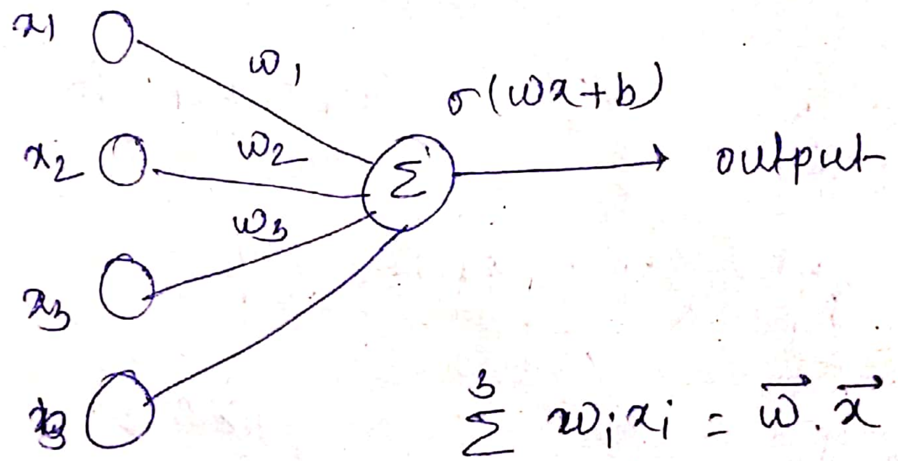
$$w(n+1) = w(n) - \eta(d(n) - y(n)) * x(n)$$

$$y[-\infty, \infty] \rightarrow [0, 1]$$

logistic function

$$f(x) = \frac{1}{1+e^{x}}$$

~~New~~ Neural Network basis : single unit



$$\sum_{i=0}^{3} w_i x_i = \vec{w} \cdot \vec{x}$$

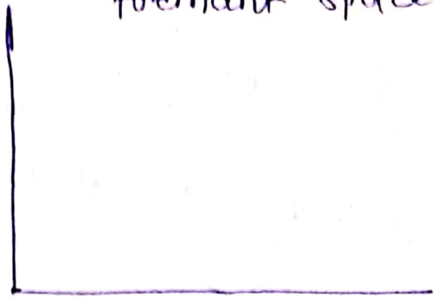
$$\sigma(w_1x_1 + w_2x_2 + w_3x_3 + b) = \sigma(Wx + b)$$

$$W \in R^{1\times 3}$$

"logistic regression as a neuron"

* vowel ~~rego~~ recognition

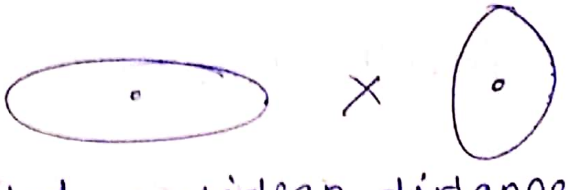
formant space of vowels



classification criteria (deterministic view)

Euclidean distance

$$x \in C_k \text{ if } (x-\mu_k)^2 \le (x-\mu_j)^2 \quad \forall j$$



weighted euclidean distance

$$d^k = \sqrt{\left(\frac{x-\mu^k}{\sigma^k}\right)^2}$$

extention to multiple features

$$d^k = \sqrt{\sum_i \left(\frac{x_i-\mu_i^k}{\sigma_i^k}\right)^2}$$

DTW : Matching sequence

lexical order

~~rigd to~~ valid word

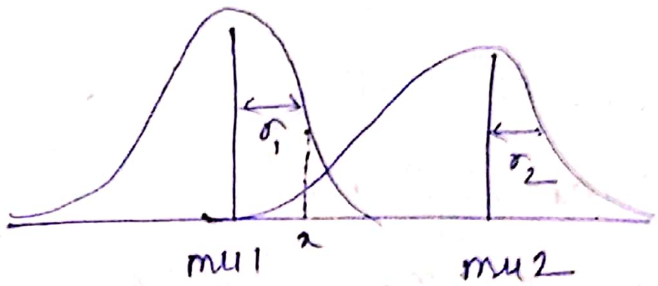
sequence of word

<u>Two class problem</u> (probabilistic view)

Normal distribution : $N(\mu, \sigma)$

Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}$$



$N(\mu_1, \sigma_1)$ $\quad\quad$ $N(\mu_2, \sigma_2)$

maximum likelihood classification criteria :

$x \in C_k$ if $p(x / N(\mu_k : \sigma_k)) \geq p(x / N(\mu_j : \sigma_j)) \quad \forall j$

$$\frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma|} \exp\left(-\frac{1}{2} \left\{ (\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \right\} \right)$$

x one speech frame

test data

Scanned by CamScanner

isolated word recognition

Ex. Name dialing

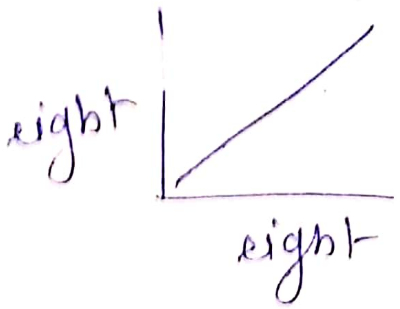
→ End-point detection errors
→ speaking rate variations
→ within word variation

* linear warping → Note

matching of feature vectors in test & reference

Greatest similarity (lesser distance)

'eight' versus 'eight' : A path diagonal exist
'eight' versus 'three' : A path diagonal doesnot exist.



* Dynamic programming

$$D(n,m) = d(n,m) + \min \begin{cases} D(n-1, m) \\ D(n-1, m-1) \\ D(n, m-1) \end{cases}$$