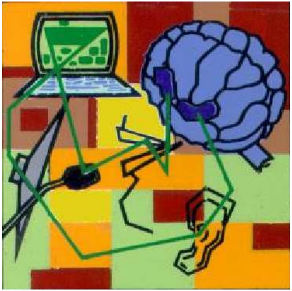
Speech Technology: Research & Applications

Samudravijaya K

C.S.Dept. Mumbai Univ. 31-DEC-12

samudravijaya@gmail.com

Tata Institute of Fundamental Research



Computer Processing of Spoken language

Language: primary mode of communication

Computer Processing of

- Written language
  - script recognition
  - fonts for display
- Spoken language
  - Speech Coding
  - Speech Recognition (ASR)
  - Speaker Recognition
  - Language / Accent / Gender / Emotion Recognition
  - Spoken Language Understanding
  - Text-to-Speech (TTS) Systems

Speech to speech translation

Speaker

Human speech production

↓

Language identification

↓

Hindi speech recognition

↓

Hindi to English
Machine translation

↓

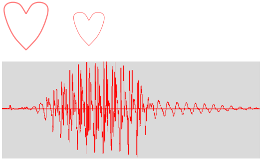
English
Text-to-speech

↓

Telephone system

↓

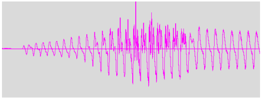
Listener

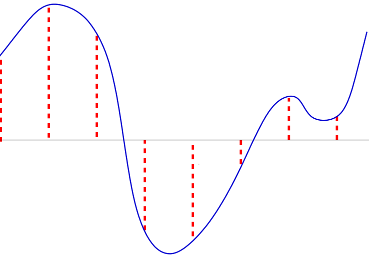


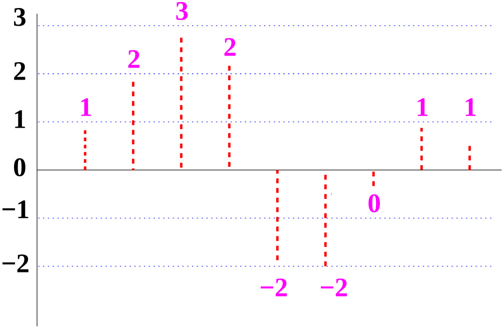
Hindi

कब आयेंगे

When will you come







A2D === Sampling + Quantization

Audio Compression / Coding

Sampling Theorem: $F_s >= F_{max}$

1sec of Music === 44100Hz X 2bytes X 2 channels $\approxeq$ 176KB
1sec of Speech === 8000Hz X 1byte $\approxeq$ 8KB

Speech Compression: Send difference between adjacent samples.
Instead of sending
100, 105, 112, 107, etc., send
100, 5, 7, -5 etc.

Audio Compression / Coding

Sampling Theorem: $F_s >= F_{max}$

1sec of Music === 44100Hz X 2bytes X 2 channels $\doteqdot$ 176KB
1sec of Speech === 8000Hz X 1byte $\doteqdot$ 8KB

Speech Compression: Send difference between adjacent samples.
Instead of sending
100, 105, 112, 107, etc., send
100, 5, 7, -5 etc.

Prediction based compression: Send difference between the actual value and a predicted value.
In the above case predicted value of 2nd sample was the value of the 1st sample (100). The actual values was 105. We transmit
5 = (actual value - predicted value).

Predicted value can be a fraction of the previous value.

Linear Predictive Coding (LPC)

Predicted value can be a fraction of the previous value.

$$\hat{x}(n) \; = \; \alpha x(n-1)$$

Send the error = difference between the actual value and a predicted value.

Linear Predictive Coding (LPC)

Predicted value can be a fraction of the previous value.

$$\hat{x}(n) \; = \; \alpha x(n-1)$$

Send the error = difference between the actual value and a predicted value.

To reduce the prediction error, predict the current sample value as a linear combination of several past sample values.
Send difference between predicted and actual sample value.

$$\widehat{s[n]} = \sum_{i=1}^{i=10} \alpha_k s[n-k]$$

$$e[n] = s[n] - \widehat{s[n]}$$

Need to send 10 real numbers (α_k) and error sequence (small values).
This requires fewer bytes than sending actual (160) samples ($s[n]$).

Linear Predictive Coding (LPC)

$$e[n] = s[n] - G[n] \sum_{i=1}^{i=10} \alpha_k s[n-k]$$

Need to send 10 real numbers (α_k) and error sequence (small values). This requires fewer bytes than sending actual samples ($s[n]$).

Speech Coding: The set of 10 real numbers can be coded as one of, say 256, representative sets.

Linear Predictive Coding (LPC)

$$e[n] = s[n] - G[n] \sum_{i=1}^{i=10} \alpha_k s[n-k]$$

Need to send 10 real numbers (α_k) and error sequence (small values). This requires fewer bytes than sending actual samples ($s[n]$).

Speech Coding: The set of 10 real numbers can be coded as one of, say 256, representative sets.

Variable Rate Coding: Change the compression rate (hence quality) depending on circumstances.

Linear Predictive Coding (LPC)

$$e[n] = s[n] - G[n] \sum_{i=1}^{i=10} \alpha_k s[n-k]$$

Need to send 10 real numbers (α_k) and error sequence (small values). This requires fewer bytes than sending actual samples ($s[n]$).

Speech Coding: The set of 10 real numbers can be coded as one of, say 256, representative sets.

Variable Rate Coding: Change the compression rate (hence quality) depending on circumstances.

GSM: LPC + RPE + LTP:
Coded 13 LPCs + Regular Pulse Excitation + Long Term Prediction

Estimation of LPCs

$$E \ = \ \sum_n \{x(n) \ - \ \hat{x}(n)\}^2$$

$$E \ = \ \sum_n \{x(n) \ - \ \sum_{k=1}^{p} \alpha_k x(n-k)\}^2$$

Minimize prediction error by setting

$$\delta E \ / \ \delta \alpha_k \ = \ 0, \qquad\qquad k = 1, 2, ...p.$$

These lead to a set of p equations in p variables, which can be solved using matrix operations.

What is ASR?

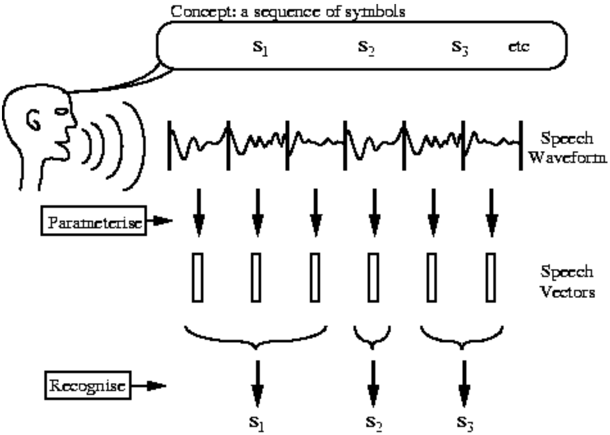


Fig. 1.1 Message Encoding/Decoding

source: HTK book

Applications of ASR

Dictation machine

Command and Control

- Speech interface to computer
- Electronic gadgets: phone, TV, VCR etc.
- Eyes and hands busy situations: Car driver, Pilot in a cockpit
- Aids to handicapped: voice operated wheel chair
- Keyword spotting
- Spoken-document-summarization
- Information retrieval: bank, travel, Telco

ipizza : multi-modal interface

"Id like to order a pizza with mushrooms and ham"

imod : movie on demand

"Action movies with Bruce Willis"

Speak4it - multimodal local business search

Show me the nearest Bank of America offices

source: http://nexus404.com/Blog/2011/10/16/ipad-2-iphone-4-siri-ports-coming-soon-developers-already-working-on-bringing-

Types of ASR

Types of speech:

- Isolated Word Recognition (IWR)
- Connected Word Recognition (CWR)
- Continuous Speech Recognition (CSR)
- Spontaneous speech
- KeyWord Spotting (KWS)

Speaker dependence:

- speaker dependent/adaptive/independent
- multi-speaker

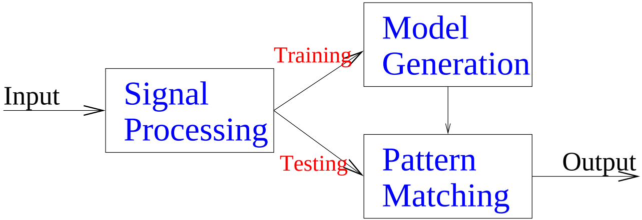
Vocabulary:

- Small ($<$ 100 words), Medium (hundreds), Large (thousands)
- Very large (tens of thousands), Out of vocabalary (OOV)

Bandwidth:

- Wideband/desktop
- Narrowband

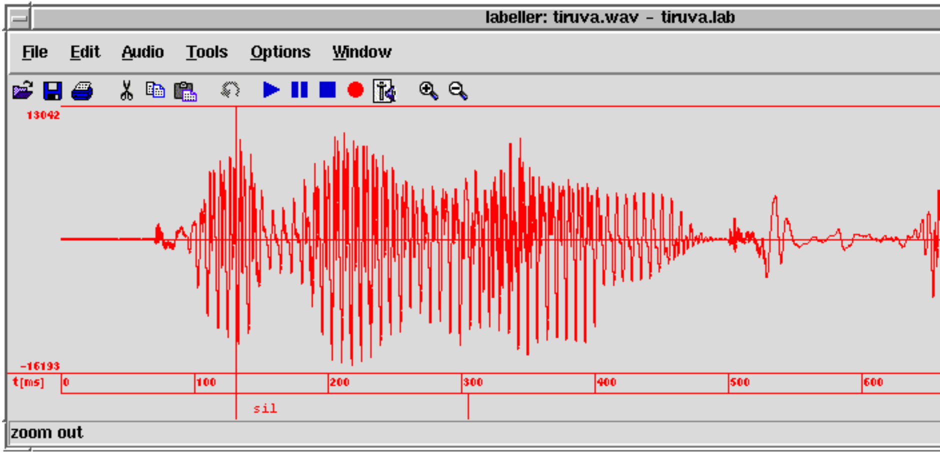
Speech Recognition is Sequential Pattern Recognition



Goal: recognise the sequence of words from time waveform of speech.

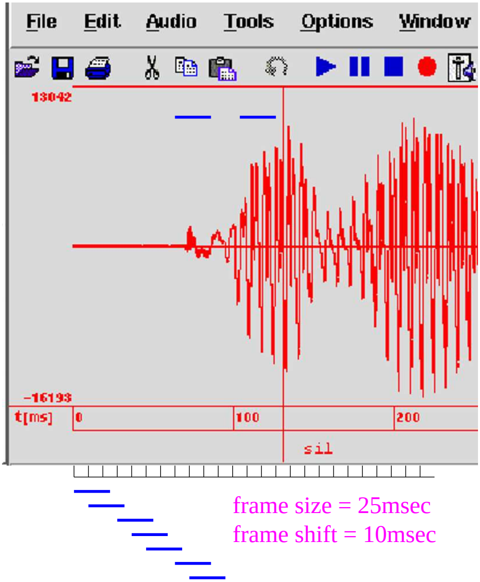
Two phases: Training (learning) and Testing (recognition)

Short-time processing of speech signal

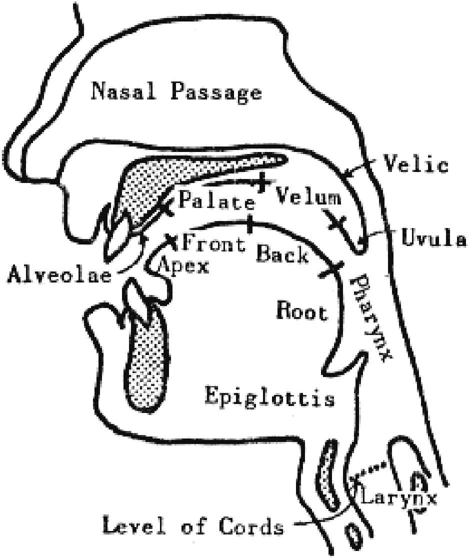


tiruva: The 3 Vowels appear similar. So, perform Spectral (Frequency) analysis

Short time speech processing



Speech Production: Articulators



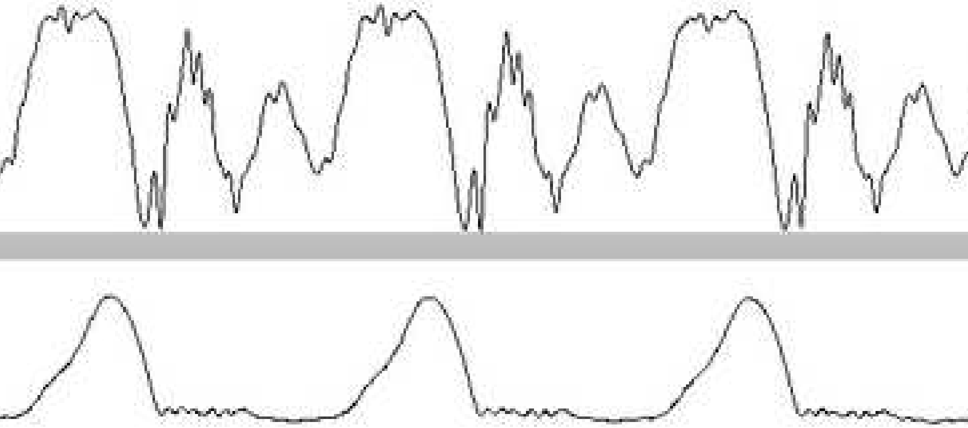
Place and Manner of Articulation

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

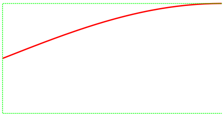
य	र	ल	व	श	ष	स	ह
y	r	l	w	sh	S	s	h

Production of voiced sounds



vowel अ

glottal vibration vocal tract speech wave



Uniform tube model

$$\nu = c/\lambda = 34000/4 * 17 = 500Hz$$

Formants $===$ poles of an all-pole model

Resonance frequency depends on dimension

Source-Filter model of speech production

glottal vibration   vocal tract   speech wave

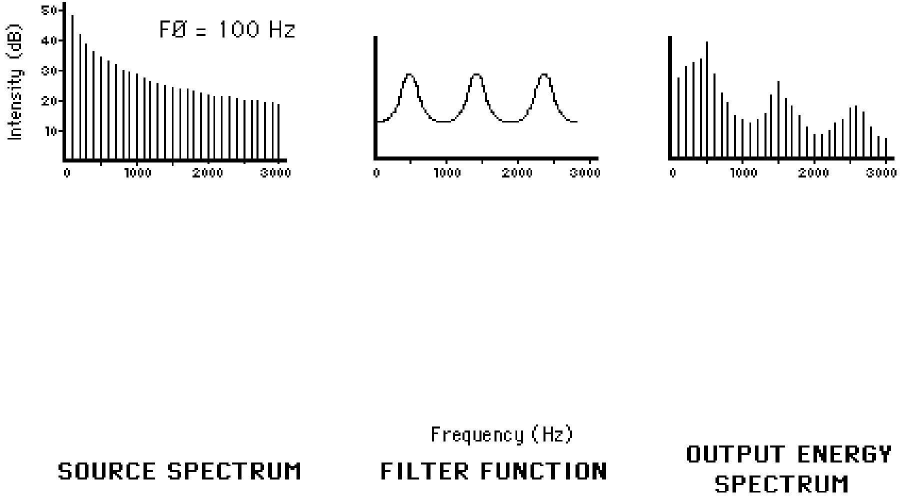
$$s(n) = e(n) * h(n)$$

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

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

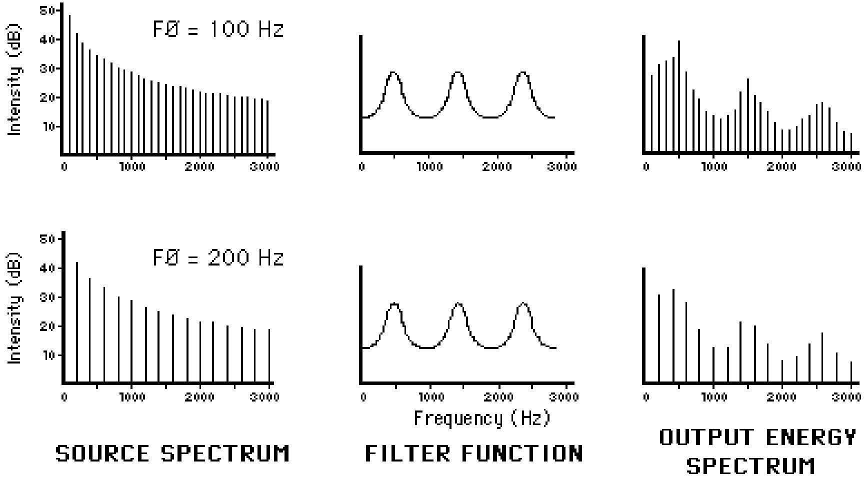
In practice, any effect that cannot be modeled by an all-pole model is called 'residual'; it represents characteristics of lip radiation in addition to excitation.

Illustration in spectral domain



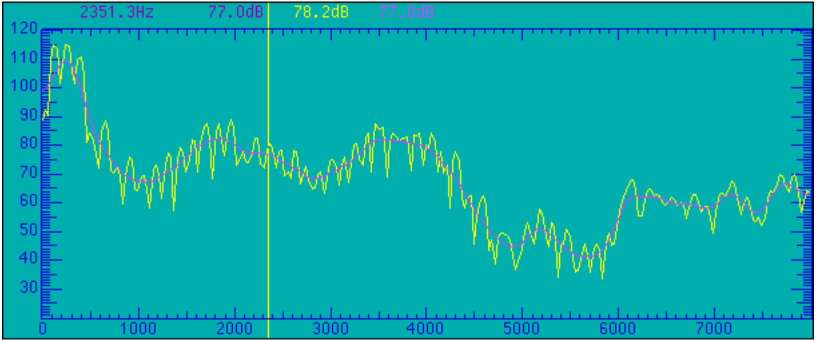
source: http://www.haskins.yale.edu/haskins/HEADS/MMSP/acoustic.html

Illustration in spectral domain



source: http://www.haskins.yale.edu/haskins/HEADS/MMSP/acoustic.html

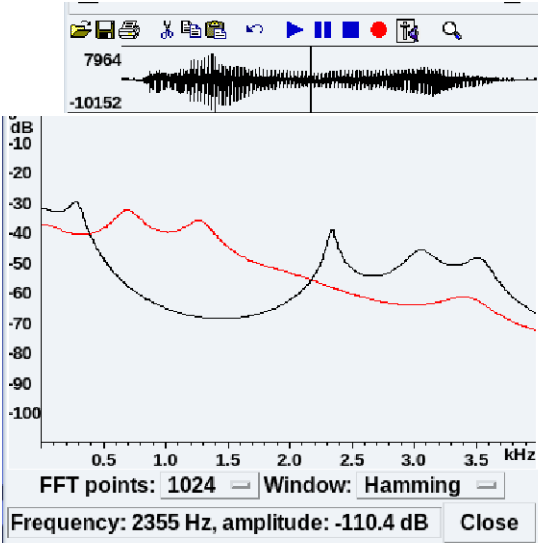
Cepstrally smoothed spectrum



Waveform Power spectrum Log spectrum Cepstrum

$$cep(q) = IFFT\{log(|S(k)| **2)\} \qquad q = 0, 1, ...N-1$$

Peaks in the spectrum are called formants.



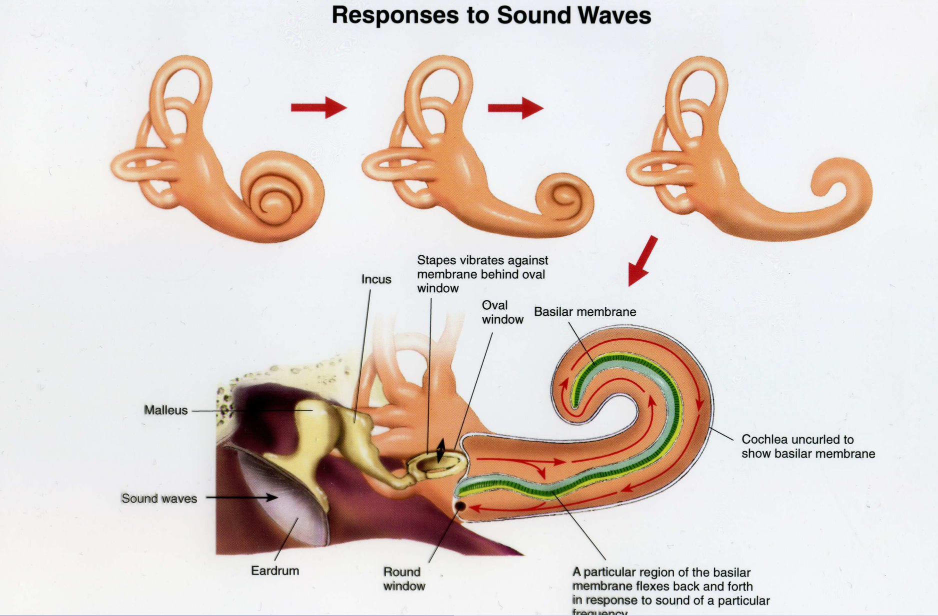
Vowel: formant frequencies (Hz) (Signatures)

/aa/: F1=700; F2=1300

/i/ : F1=300; F2=2300

/e/ : F1=350; F2=2100

Hint from Biology



Frequency Analysis by Cochlea

Cochlear response: animation (source: https://courses.washington.edu/psy222/auditory%20demos/)

Basilar membrane: Bark/mel scale

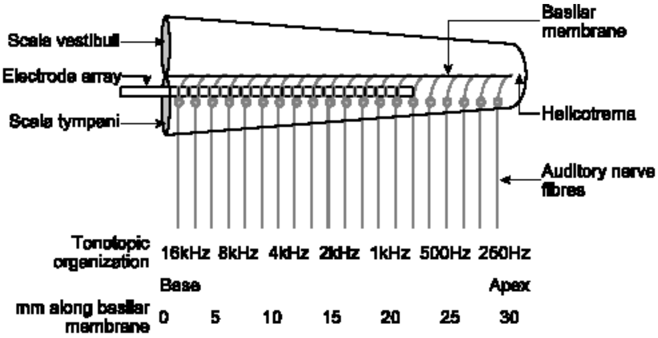
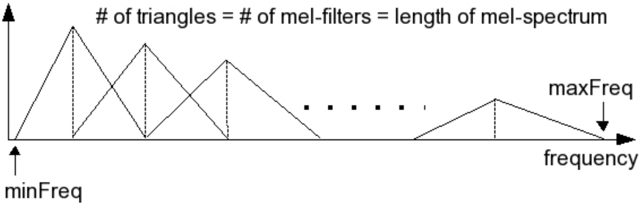


Figure 1.1. A simplified unrolled representation of the cochlea showing the auditory nerve fibres, the tonotopic organization of these nerve fibres and an intracochlear electrode array in the scala tympani.

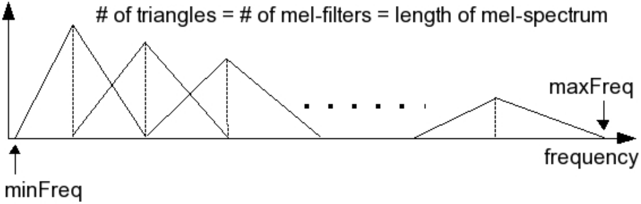
Critical band phonomenon

Non-linearities along amplitude and frequency

Filter-bank analysis: MFCC

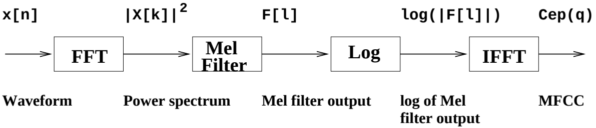


Filter-bank analysis: MFCC



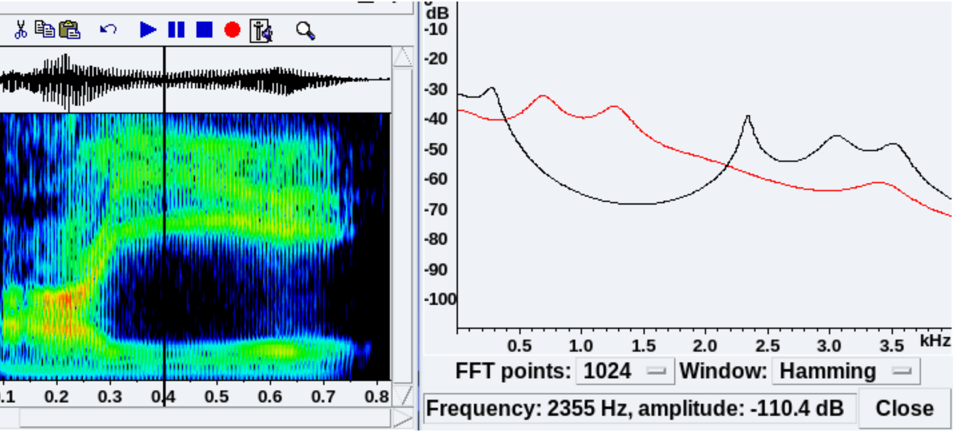
$$B(m) = \sum_{k=lo(m)}^{hi(m)} |X(k)|^2$$

$$cep(q) = IFFT\{log(|B(m)|^2)\} \qquad q = 0, 1, \ldots N$$



Mel Frequency Cepstral Coefficients: **MFCC**

Temporal Variation of spectrum: Spectrogram



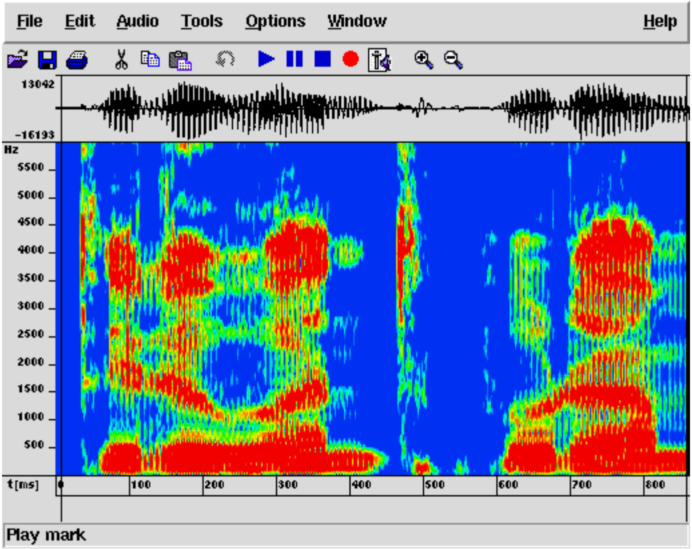
Vowel: formant frequencies (Hz) (Signatures)

/aa/: F1=700; F2=1300

/i/ : F1=300; F2=2300

/e/ : F1=350; F2=2100

Spectrogram



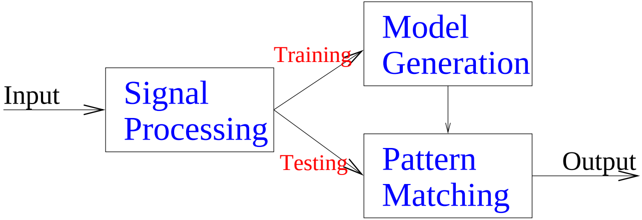
Formant: frequency of resonance: F1, F2, F3, ...

Speech: a dynamic signal: Slope and curvature of formant trajectories.

Speech Signal Processing (Feature Extraction)

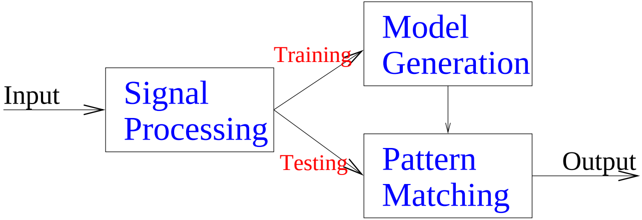
- Digitisation of analog speech signal
- Blocking signal into frames
- FFT $\rightarrow$ mel filter $\rightarrow$ log $\rightarrow$ IFFT $\Rightarrow$ MFCC
- Slope and curvature
- Sequence of feature vectors : $x_1, x_2, \ldots x_T$

Static Pattern Recognition



Signal Processing $\Rightarrow$ Sequence of feature vectors

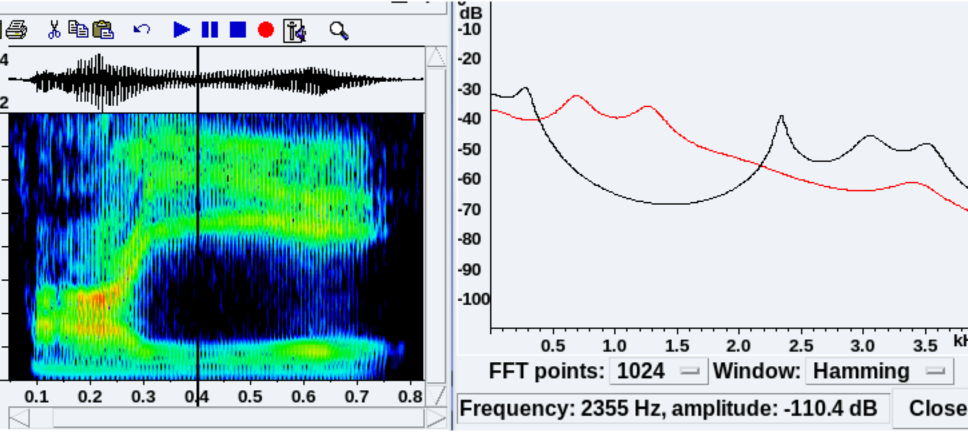
Static Pattern Recognition



Signal Processing ⇒ Sequence of feature vectors

Pattern Recognition

Illustration: Vowel recognition with the first 2 Formant frequencies as features



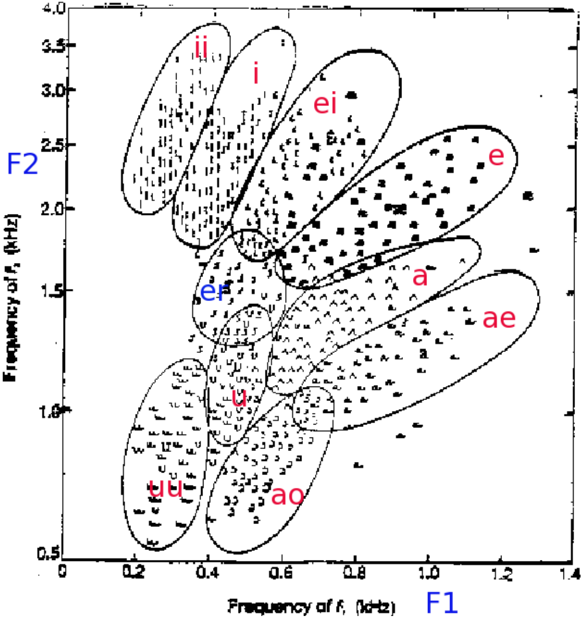
Vowel: formant frequencies (Hz) (Signatures)

/aa/: F1=700; F2=1300

/i/ : F1=300; F2=2300

/e/ : F1=350; F2=2100

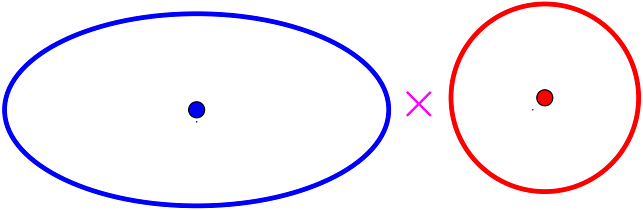
Formant space of vowels



Classification criterion

* *E*uclidean Distance

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



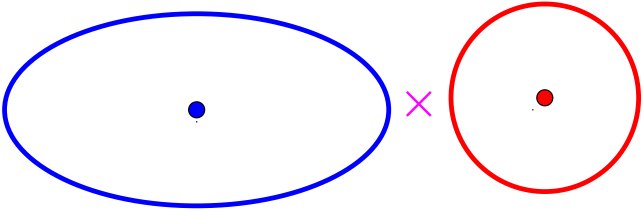
* Weighted Euclidean distance

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

Classification criterion

* *E*uclidean Distance

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



* Weighted Euclidean distance

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

* Extension to multiple features

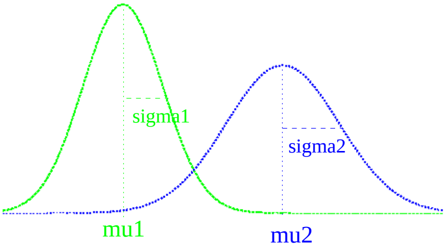
$$d = \sum_i \left(\frac{\mathbf{x_i} - \mu_\mathbf{i}^\mathbf{k}}{\sigma_i}\right)^2$$

Probabilistic models

Two class problem

Normal Distribution: $N(\mu;\sigma)$

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

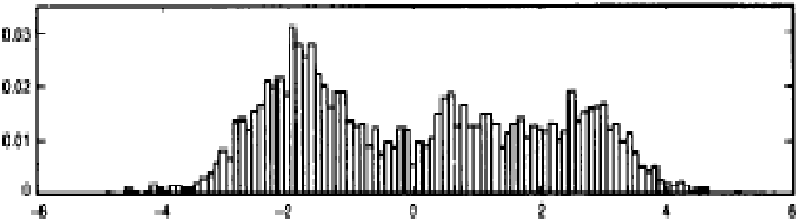


Maximum Likelihood classification criterion:

$$x \in C_k \quad \text{if } p(x|N(\mu_k;\sigma_k)) \;\geq\; p(x|N(\mu_j;\sigma_j)) \quad \forall j$$

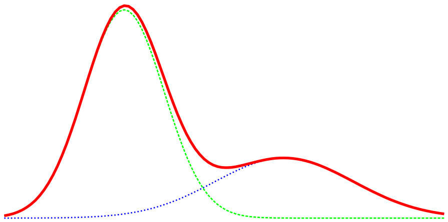
Refer to vowel F1-F2 diagram

Features need not follow Normal Distribution



Histogram

Gaussian Mixture Model(GMM)



$$p(x|GMM(k)) = \alpha p(x : N[\mu_1; \sigma_1]) \quad + (1-\alpha)\ p(x : N[\mu_2; \sigma_2])$$

Maximum Likelihood classification criterion for GMM case:

$$x \in C_k \quad \text{if } p(x|GMM(k)) \geq p(x|GMM(j)) \quad \forall j$$

Extension to Multi-dimensional space

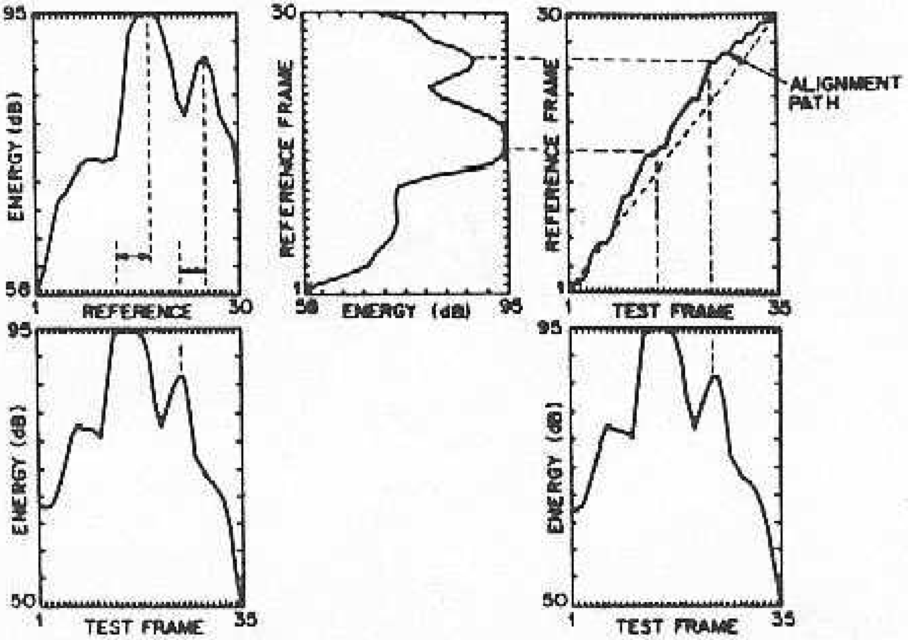
Classification of Temporal patterns

Isolated Word Recognition:
Example: name dialling

Match a *s*equence of test feature vectors $x_1, x_2, \ldots, x_N$
with a *s*equence of reference feature vectors $r_1, r_2, \ldots, r_M$
Reasons for $N \neq M$

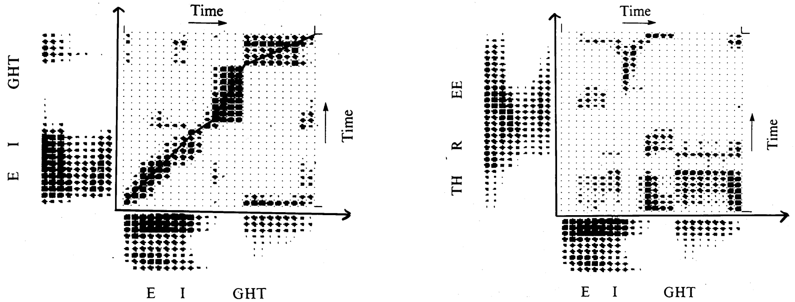
- End-point detection errors
- speaking rate variations
- Within word variations

Linear vs Non-linear Time-warping



From: Fundamentals of Speech Recognition, L.Rabiner and B.H.Juang

Optimal alignment path



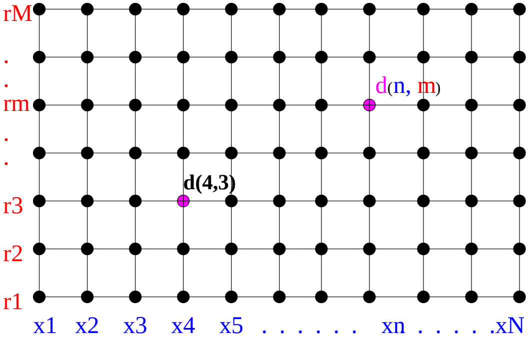
From: Holmes book

Bigger the dark blob, greater the similarity (lesser distance).

“eight” versus “eight”: A path along diagonal exists

“eight” versus “three”: A path along diagonal does not exist.

Dynamic Programming

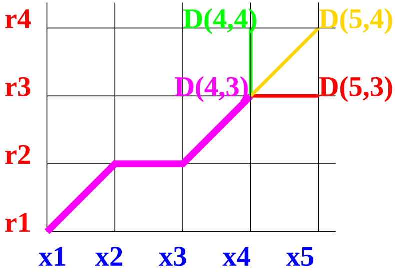


Test feature vector sequence

Goal: To find the optimal alignment path from the grid point $(1,1)$ to the grid point (N, M). There are exponential number (M^N) of paths. In order to reduce the number of computations from exponential to linear, we use the Dynamic Programming whose foundation is the "Principle of Optimality".

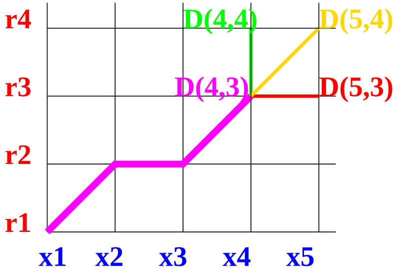
Principle of optimality: The best path from $(1,1)$ to any given point on the grid is independent of what happens beyond that point.

So, if two paths share a partial path starting from $(1,1)$, the cost of this shared partial path need to be computed only once and stored in a table for later use.



Principle of optimality: The best path from $(1, 1)$ to any given point on the grid is independent of what happens beyond that point.

So, if two paths share a partial path starting from $(1, 1)$, the cost of this shared partial path need to be computed only once and stored in a table for later use.



DP Algorithm: Define

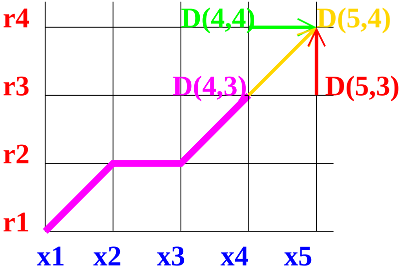
$d(n, m)$: the **local** distance between the n^{th} test frame and m^{th} reference frame.

$D(n, m)$: the **accumulated** distance of the optimal path starting from the grid point $(1, 1)$ and ending at the grid point (n, m): cost of shared path.

Dynamic Time Warping

Applying the *P*rinciple of optimality, $D(n,m)$ is the sum of the local cost, and the cost of cheapest path to it

$$\mathbf{D(5,4) = d(5,4) + min} \begin{cases} \mathbf{D(4,4)} \\ \mathbf{D(4,3)} \\ \mathbf{D(5,3)} \end{cases}$$

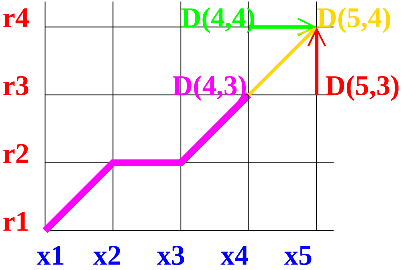


$$D(n,m) = d(n,m) + min \begin{cases} D(n-1,m) \\ D(n-1,m-1) \\ D(n,m-1) \end{cases}$$

Dynamic Time Warping

Applying the *P*rinciple of optimality, $D(n,m)$ is the sum of the local cost, and the cost of cheapest path to it

$$\mathbf{D(5,4) = d(5,4) + min} \begin{cases} \mathbf{D(4,4)} \\ \mathbf{D(4,3)} \\ \mathbf{D(5,3)} \end{cases}$$

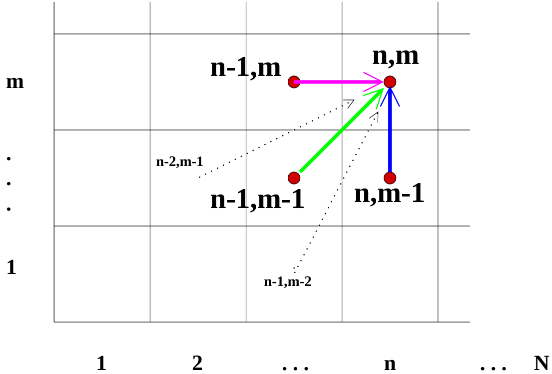


$$D(n,m) = d(n,m) + min \begin{cases} D(n-1,m) \\ D(n-1,m-1) \\ D(n,m-1) \end{cases}$$

$*$ Compute $D(n,m)$ for each "allowed" pair of (n,m).
Remember the "best" predecessor point.

$*$ $D(N,M)$ is the cost of the optimal path.

$*$ From (N,M), start backtracing to identify the optimal path.



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

$*$ Compute $D(n,m)$ for each “allowed” pair of (n,m).
Remember the “best” predecessor point.

$*$ $D(N,M)$ is the cost of the optimal path.

$*$ From (N,M), start backtracing to identify the optimal path.

Global constraints: left- and down-paths are prohibited.

Local constraints: path $(n,m-1) \rightarrow (n,m)$ not allowed.

Spell checking: Application of Dynamic Programming

Reference (correct spelling)

	p	p	a	t	a	r	r	n
n	1	1	2			1	1	0
r	1	1	2	1	2	0	0	1
e	2	2	1	2	1	2	2	2
t	1	1	2	0	2	1	1	
t	1	1	2	0	2	1	1	
a	2	2	0	2	0	2		
p	0	0	2		2	1		

Test sequence (just typed in text)

d(V,C)=2

d(V1,V2)=1

d(C1,C2)=1

n	1 / 8	1	2			1 / 2	1 / 2	0 / 1
r	1 / 7	1	2	1	2 / 3	0 / 1	0 / 1	1 / 2
e	2 / 6	2	1	2 / 2	1 / 1	2 / 3	2	2
t	1 / 4	1 / 4	2 / 4	0 / 0	2 / 2	1	1	
t	1 / 3	1 / 3	2 / 2	0 / 0	2 / 2	1	1	
a	2 / 2	2 / 2	0 / 0	2 / 2	0	2		
p	0 / 0	0 / 0	2 / 2	2 / 4	2	1		
	p	p	a	t	a	r	r	n

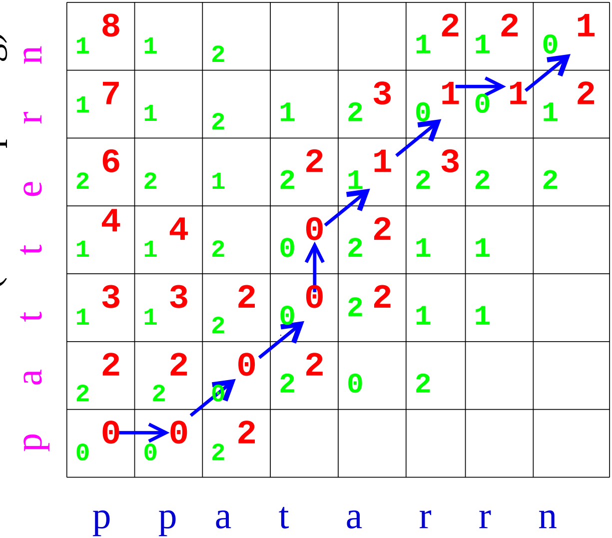
Test sequence (just typed in text)

$d(v,c)=2$

$d(v_1,v_2)=1$

$d(c_1,c_2)=1$

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

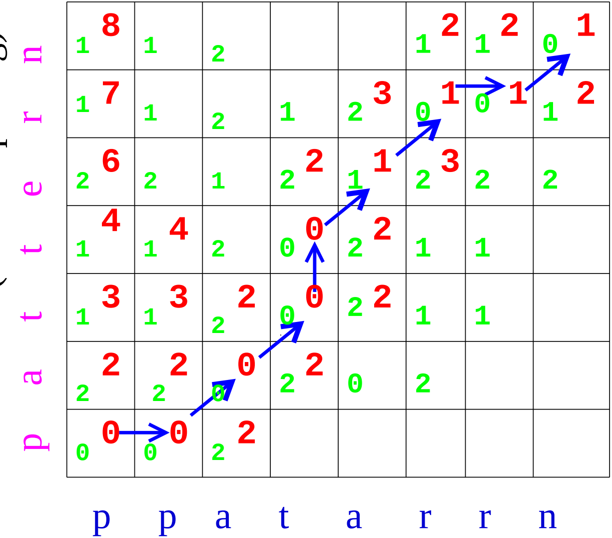


d(v,c)=2

d(v1,v2)=1

d(c1,c2)=1

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



$d(v,c)=2$

$d(v1,v2)=1$

$d(c1,c2)=1$

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

Reference template generation: average frames belonging same speech sound.

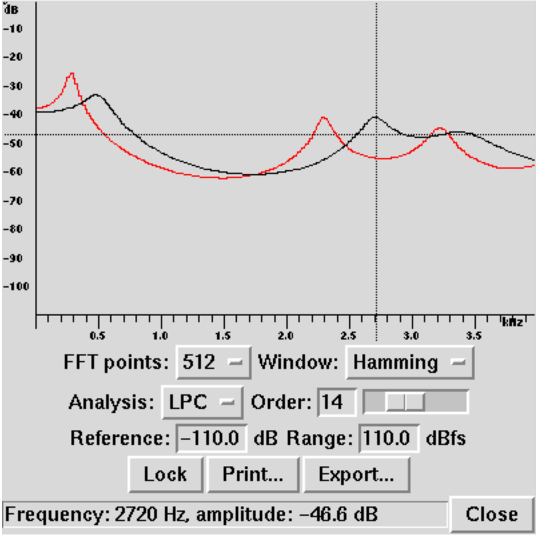
Why speech recognition is difficult?

maahitNaahi.wav

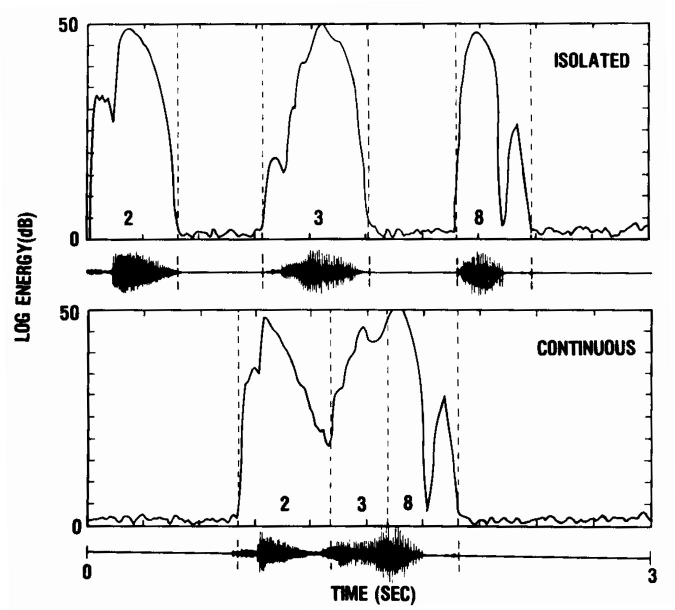
Sources of variabilities

- Speaker specific: physiological, emotional, cultural
- Continuous signal: no well defined boundaries between linguistic units
- Ambience: noise, Lombard effect, room acoustics
- Channel: additive/convolutional noise, compression
- Transducer: omni/uni-directional, carbon/electret mic
- Phonetic context

Spectra of the vowel 'i' in word "pin" spoken by male and female speakers



No well defined boundaries between linguistic units



Diversity of transduction characteristics of microphones

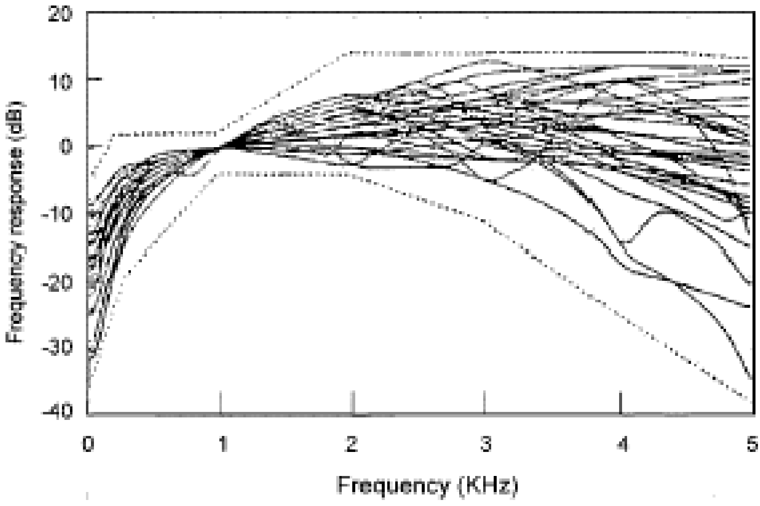
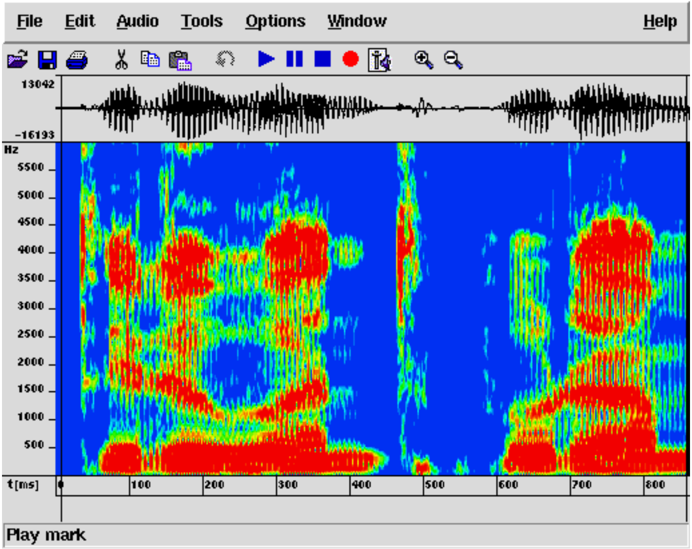


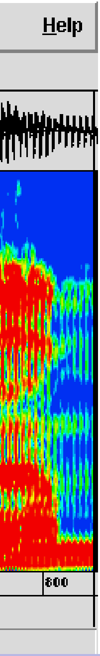
Fig. 6. Diversity of transducer characteristics in telephone set [25].

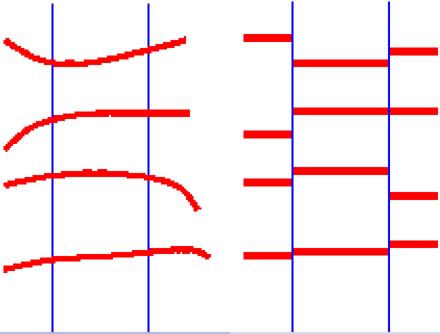
Spectrogram of thiruvananthapuram



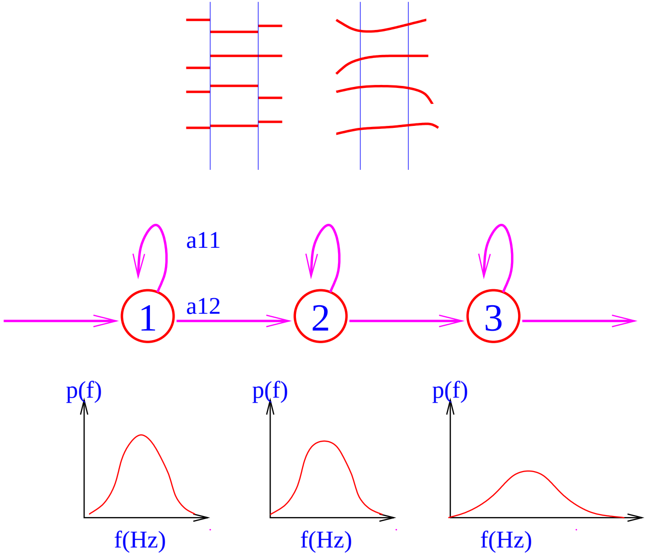
t i r u w a n th p u r a m

Formant trajectories





hidden Markov model (HMM)



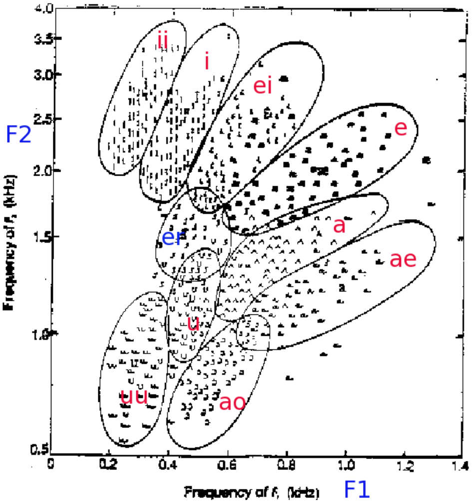
Parameters of a HMM: A, B, π

doubly stochastic model

3 problems in HMM

- How to compute the likelihood of a trained model generating a test observation sequence?
  Solution: forward algorithm (recursion used)
- How to find the optimal state sequence?
  Solution: Viterbi algorithm (similar to DTW)
- How to estimate the parameters of the model: $\lambda = (A, B, \pi)$?
  Solution: Forward-backward algorithm.

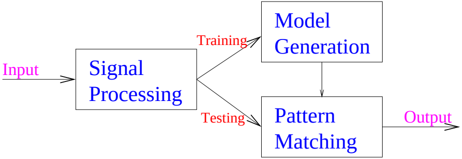
What is hidden in hidden Markov model?

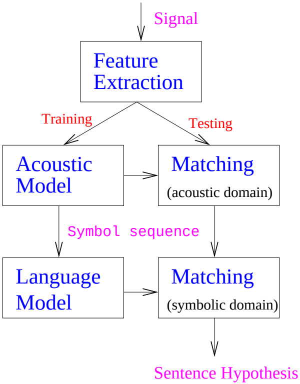


A given (F1, F2) point can be perceived as either V1 or V2 depending on speaker and context.

DTW Vs HMM

DTW involves matching the test feature sequence with reference feature sequences (reference templates) of different words and choosing the word corresponding to the minimum distance. In order to capture variability of speech, one can generate a composite reference template by averaging time-aligned feature sequences of repetitions of the same word. This improves the representation as it incorporates the first order statistics. HMM can be seen as an extension of this approach that incorporates second order statistics as well.





How a spoken sentence is recognized?

Knowledge sources

Phone sequence/phone hypothesis lattice

==> Sentence hypothesis

Lexicon

kal

lak

Lexical knowledge: Can a word begin with /ng/?

NGHALCHAWM CAMP

Letter to sound rules

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	स	ह
y	r	l	w	sh	S	s	h

Pronunciation dictionary:

kalam vs kamal

karnaa, pahale, Bhaartiya

Knowledge sources

Phone sequence/phone hypothesis lattice

$==>$ Sentence hypothesis

Lexicon

man

mna

Syntax

Some man brought the apple.

Apple the brought man some.

Knowledge sources

Phone sequence/phone hypothesis lattice

==> Sentence hypothesis

Lexicon

man

mna

Syntax

Some man brought the apple.

Apple the brought man some.

Semantics

Time flies like an arrow

Fruit flies like banana

Knowledge sources

Phone sequence/phone hypothesis lattice

==> Sentence hypothesis

Lexicon

man

mna

Syntax

Some man brought the apple.

Apple the brought man some.

Semantics

Time flies like an arrow

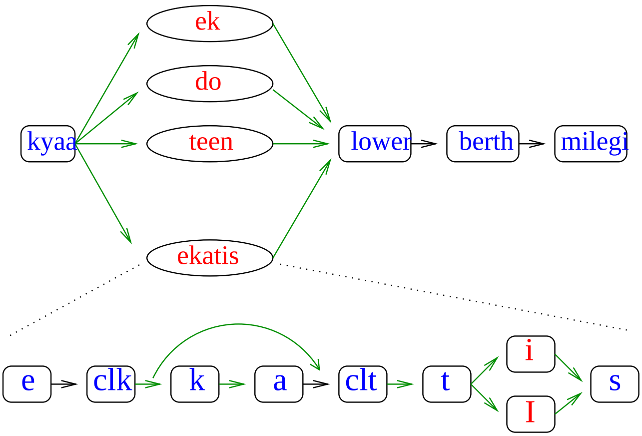
Fruit flies like banana

Pragmatics

Turn left for the nearest chemist.

Because the closest one (on the right) is closed today.

Word transition net



e clk k a clt t I s
e clk k a clt t i s

e clk clt t I s
e clk clt t i s

Pronunciation variations

Incorporation of syntax

Network grammar integrates of syntax, semantics and domain knowledge.

[क्या] Trainname (का | मे) [Digit] (रिज़र्वैशन | Class का टिकट) Aaj के लिए Milegaa [क्या]?;

Incorporation of syntax

Network grammar integrates of syntax, semantics and domain knowledge.

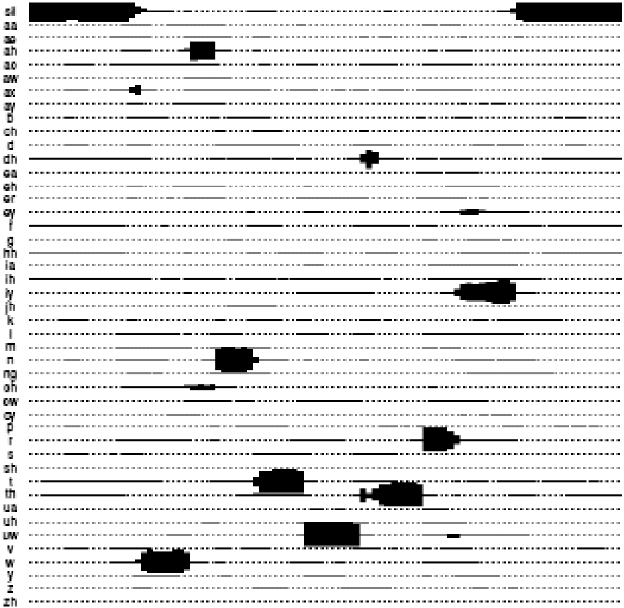
[क्या] Trainname (का | मे) [Digit] (रिज़र्वेशन | Class का टिकट) Aaj के लिए Milegaa [क्या]?;

Statistical model: Probability of word sequences

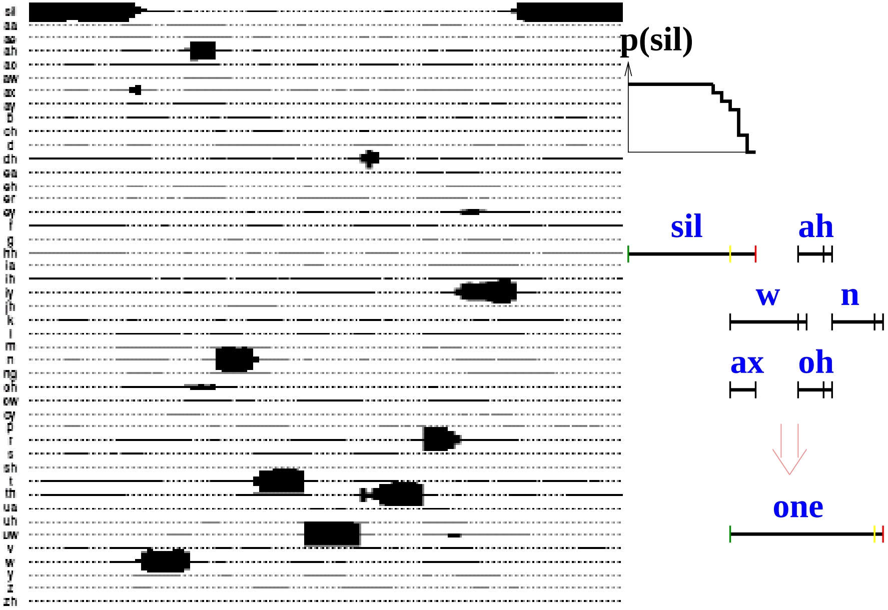
bigram: $p(w_n | w_{n-1})$

The concept can be extended to sequence of n words: n-grams

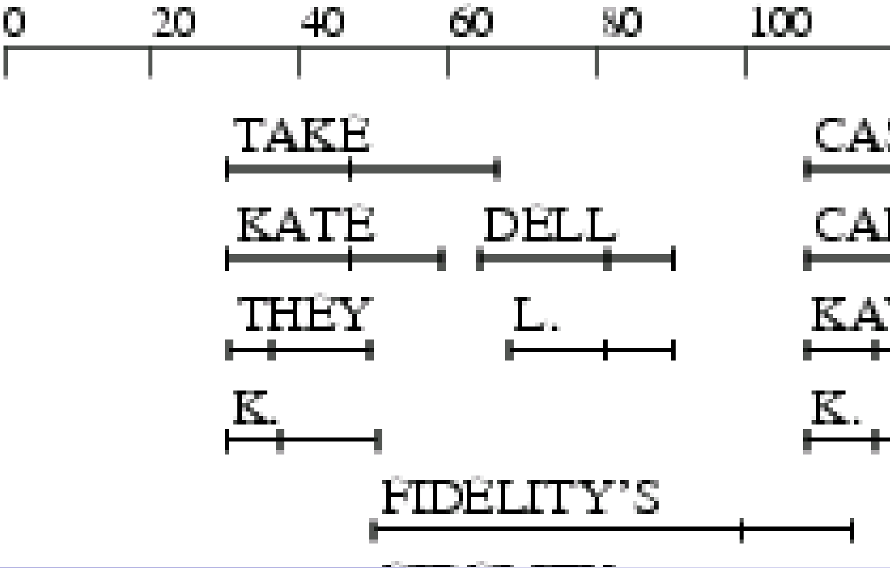
Probabilities of phones at various time instants



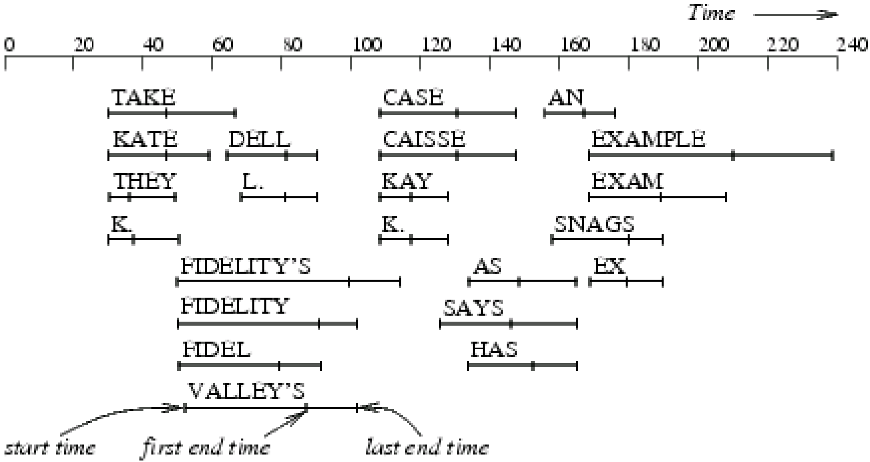
Probabilities of phones at various time instants



Lattice of phone hypotheses → lattice of word hypotheses



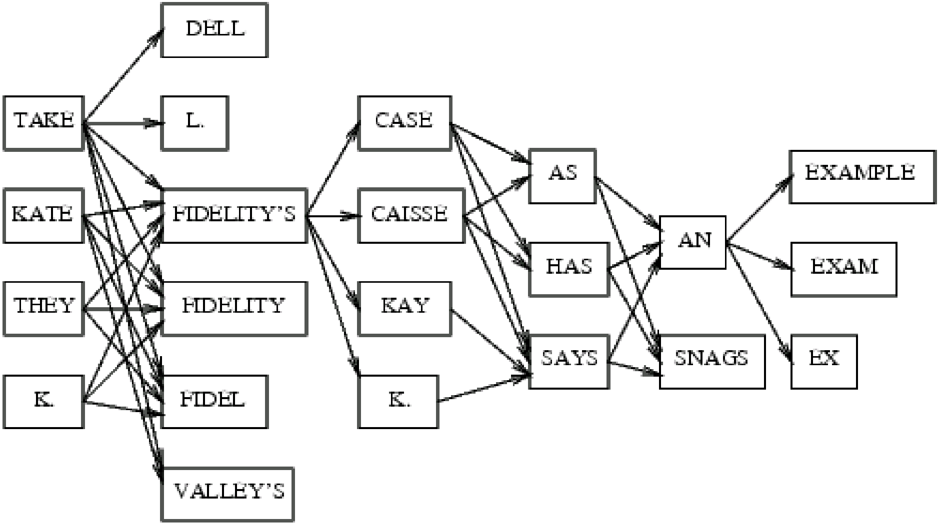
Word hypotheses at various time instants



Take Fidelity's case as an example

Source: "Efficient algorithms for Speech Recognition", M.K.Ravishankar, PhD thesis: CMU-CS-96-143

Word Lattice as a Directed Acyclic Graph



Automatic Speaker Recognition

Speech Signal contains a variety of information

Speaker Recognition is **Complimentary** to Speech Recognition

Other person identification methods

- Non-biometric: password, PIN, key
- Biometric
  - ∗ Finger print, Retina scan
  - ∗ Speech: Advantage: remote usage

- Speaker **Identification**
  - ∗ Closed set : N outcomes
  - ∗ Open set : N+1 outcomes
- Speaker **Verification**
  - ∗ binary decision
- Speaker **Tracking**
- Speaker **Segmentation**

Block diagram of Speaker Verification

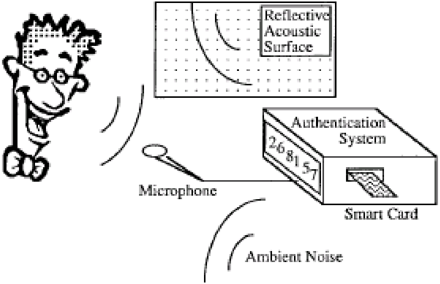
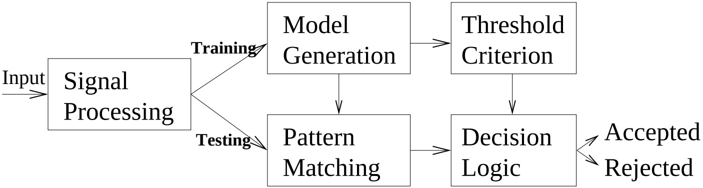


Fig. 2. Typical speaker-verification setup.



Speech : phoneme ⇔ resonance

Speaker: voice ⇔ ???

Supra-segmental features

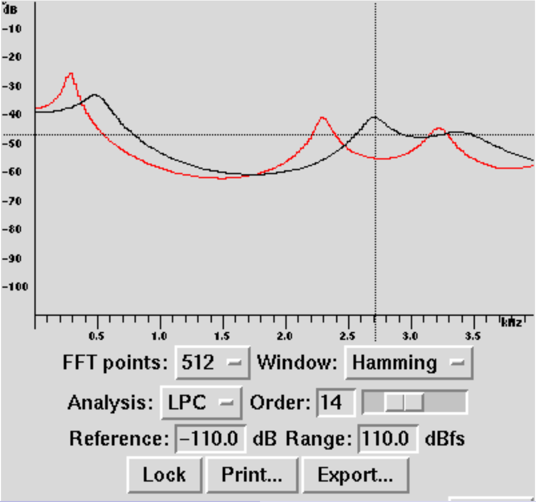
- Pitch (F0)
- Rhythm
  - Speaking Rate (duration)
  - Stress (Amplitude contour)
- Long-term statistics
  – average spectrum/cepstrum
  – OK for long utterances
- Dialect, ideolect

Voice Quality

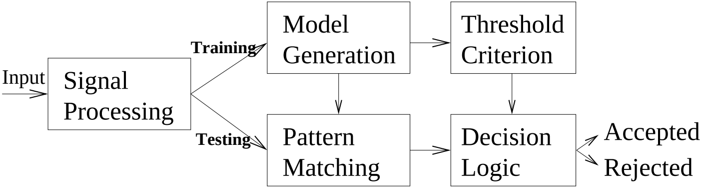
Pleasing, hoarse, resonant, breathy, nasal

Source-filter model

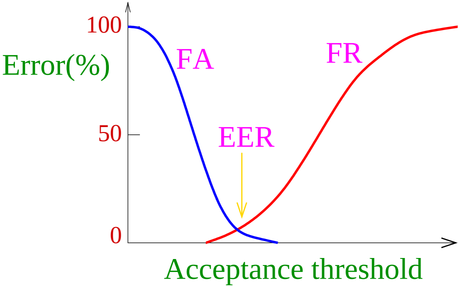
Segmental features
Linear Prediction (LP) or mel filter cepstral coefficients (MFCC);



Performance Measures



VA : Valid Acceptance
FA : False Acceptance
VR : Valid Rejection
FR : False Rejection



Equal Error Rate: EER : FA = FR

Modes of verification

Text independent

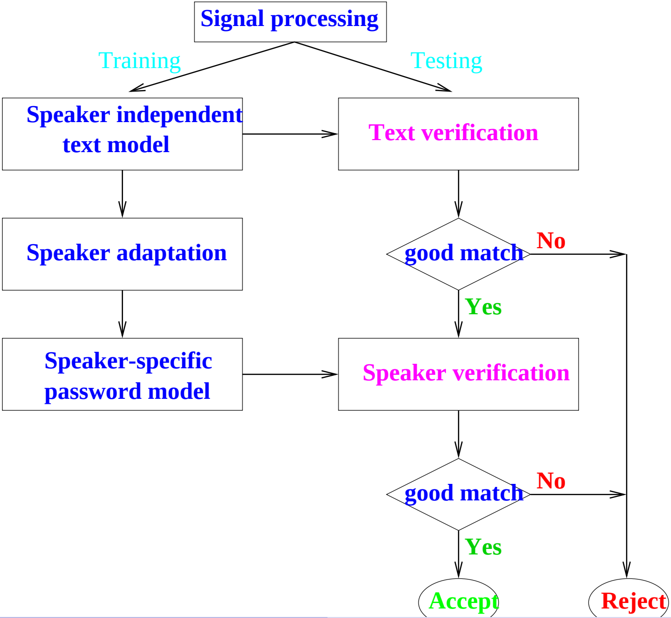
- Sequence of phonemes not known
- No control over speech
- No repetition of keywords
- useful for surviellance/forensic
- Models
  - Gaussian Mixture Model (GMM)
  - Ergodic Hidden Markov Model (HMM)
  - Support Vector Machines (SVM)

Modes of verification

Text dependent (spoken password)

- Access to secure place
- Co-operative user
- Temporal aspect of speech is relevant
- Dynamic Time Warping (DTW) or HMM can be used
- Chances of replay attack

Text prompted speaker verification



Language Identification

Applications

- Call centres
- Multi-language translation system
- Surveillance

Approaches

- Explicit identification (phone recognition)
  - acoustic score
  - frequencies of linguistic units
  - joint-likelihood score
- Implicit identification
  - vector quantisation
  - GMM

Speech Output Systems

- Limited text (Voice Response) systems
  Compressed or encoded (LPC) speech
  Applications:
  - speaking toys
  - warning systems
  - railway announcements
- Unrestricted Text (Text-to-Speech) Systems
  - text to ‘phoneme’ conversion
  - phoneme to speech conversion
  - application of prosody

Text Analysis

Input : text (Hindi or Indian English)
Output: phoneme symbols and stress markers

Phoneme repertoire: a superset of Hindi and Indian English
Stress markers: full stop, semi-colon, comma, interrogation, exclamation and end of word symbols.
Classification of text categories: Date, time, currency, alphanumerics, acronym, abbreviations, special characters, numbers(with decimal point), **text words**

Phonetic dictionary
Morphological analysis
Letter-to-phoneme rules

Phoneme-to-speech conversion

Methodologies

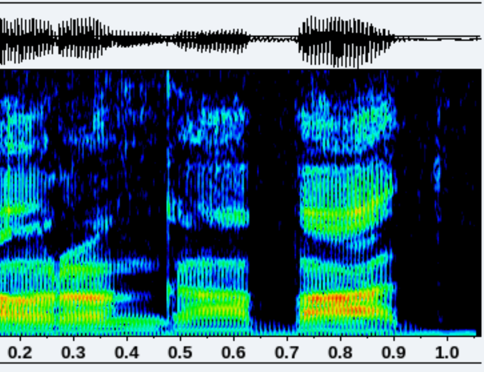
- cancatenation-based
- Model based
  - articulatory
  - formant
  - HMM

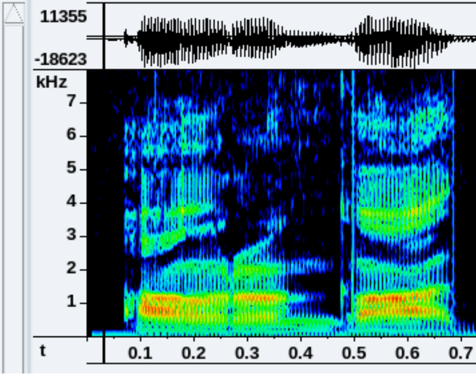
Waveform cancatenation method

Sub tasks:

- selection of basic units
- generation
- concatenation

Problems with simple-minded concatenation





Aurangabad

Aurangabad with "a" exchanged

Bibliography

- "Tutorials" at http://speech.tifr.res.in/tutorials/index.html
- "Speech and Speaker Recognition", http://speech.tifr.res.in/c̃hief/publ/03iwtdil_spSpkrReco.pdf
- "A Tutorial on Text-Independent speaker Verification", F.Bimbot et al., EURASIP J. on Appl. Sig. Processing, 4, 2004, pp. 430-451.
- "Speaker Recognition: A tutorial", J P Campbell Jr., Proc. IEEE, **85**(9), pp. 1437-1462, 1997.
- "Speaker recognition and its commercial and forensic applications", special issue of Speech Communication, vol. 31, Issue 2-3, June 2000
- "Support vector machines for speaker and language recognition" W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, P.A. Torres-Carrasquillo, Computer Speech and Language, (20)2006, pp. 210-229.

Introductory books

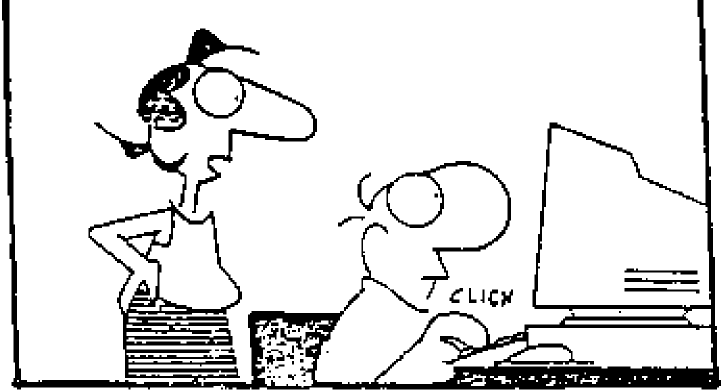
- *Speech and Audio Processing*, Shaila D. Apte, February 2012, Wiley India; Rs. 429; ISBN 978-81-265-3408-1 http://www.wileyindia.com/wileyprecise/index.php?page_id=bookdetails
- *Speech Communication: Human and Machine*, 2nd edition, Douglas O'Shaughnessy; November 1999, Wiley-IEEE Press; ISBN: 0-7803-3449-3
- *Discrete-Time Processing of Speech Signals*, Deller, Hansen, Proakis IEEE Press
- *Speech recognition by machine* W.A. Ainsworth, London: Peregrinus for the Institution of Electrical Engineers, c1988
- *Speech Coding and Synthesis*, W.B. Kleijn and K.K. Paliwal (Eds.), Elsevier, Amsterdam, 1995.

Advanced books

- *Spoken Language Processing : A Guide to Theory, Algorithm and System Development* Xuedong Huang, Raj Reddy
- *Fundamentals of Speech Recognition*, Lawrence Rabiner & Biing-Hwang Juang, Englewood Cliffs NJ: PTR Prentice Hall (Signal Processing Series), c1993, ISBN 0-13-015157-2
- *Hidden Markov models for speech recognition* X.D. Huang, Y. Ariki, M.A. Jack. Edinburgh: Edinburgh University Press, c1990.
- *Statistical methods for speech recognition*, F.Jelinek, The MIT Press, Cambridge, MA., 1998.

Software

- *Speech and Audio Processing*, Shaila D. Apte, February 2012, Wiley India; Rs. 429; ISBN 978-81-265-3408-1 http://www.wileyindia.com/wileyprecise/index.php?page_id=bookdetails DVD accompanying the book has matlab code for many speech processing applications.
- *Matlab Audio Processing*, http://www.ee.columbia.edu/ dpwe/resources/matlab; MFCC, DTW.
- *VOICEBOX: Speech Processing Toolbox for MATLAB*, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html; file I/O, LPC, MFCC
- *HTK - Hidden Markov Model Toolkit - Speech Recognition toolkit*, http://htk.eng.cam.ac.uk; linux / MS Windows
- *Sphinx: Open Source Speech Recognition Engine*, http://cmusphinx.sourceforge.net/html/cmusphinx.php; linux / MS Windows .
- *festival/festvox, software for concatenative speech synthesis*



"What good is a faster computer, faster modem and faster printer if you're still using the same old slow fingers?"

Times of India, 19-OCT-1998