DT2112
Speech Recognition by Computers

Giampiero Salvi

KTH/CSC/TMH giampi@kth.se

VT 2014

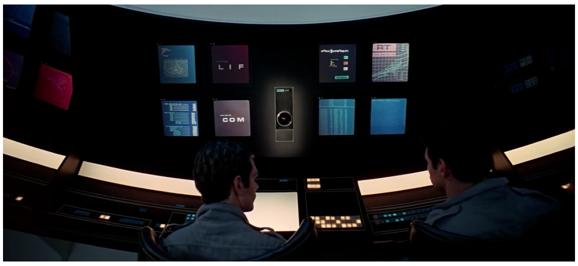
Notes

Motivation

- Natural way of communication (No training needed)
- Leaves hands and eyes free (Good for functionally disabled)
- Effective (Higher data rate than typing)
- Can be transmitted/received inexpensively (phones)

Notes

A dream of Artificial Intelligence

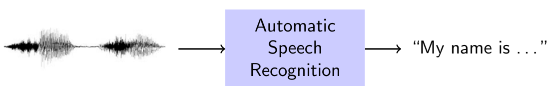


2001: A space odyssey (1968)

Notes

The ASR Scope

Convert speech into text



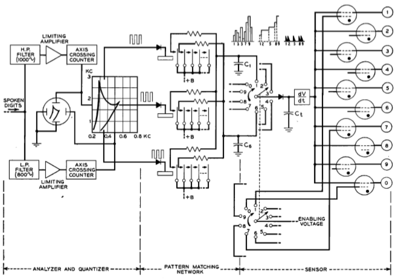
Not considered here:

- non-verbal signals
- prosody
- multi-modal interaction

Notes

A very long endeavour

1952, Bell laboratories, isolated digit recognition, single speaker, hardware based [2]



[2] K. H. Davis, R. Biddulph, and S. Balashek. "Automatic Recognition of Spoken Digits". In: *JASA* 24.6 (1952), pp. 637–642

Notes

An underestimated challenge

for 60 years many bold announcements

Notes

Applications today

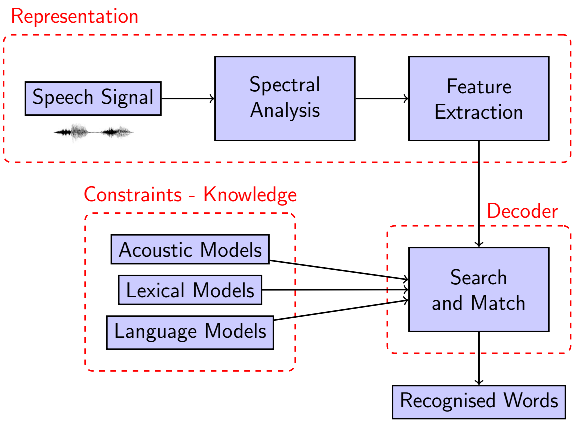
Call centers:
- traffic information
- time-tables
- booking...

Accessibility
- Dictation
- hand-free control (TV, video, telephone)

Smart phones
- Siri, Android...

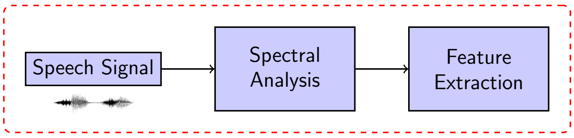
Notes

Components of ASR System



Notes

Speech Signal Representations



Goals:

- disregard irrelevant information
- optimise relevant information for modelling

Means:

- try to model essential aspects of speech production
- imitate auditory processes
- consider properties of statistical modelling

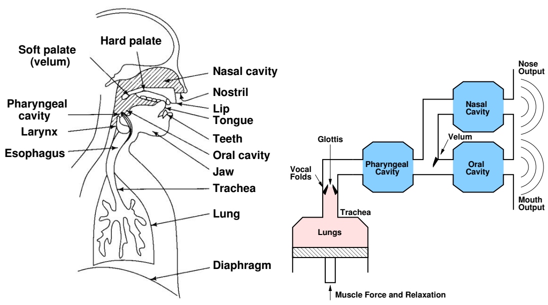
Notes

Examples of Speech Sounds

http://www.speech.kth.se/wavesurfer/

Notes

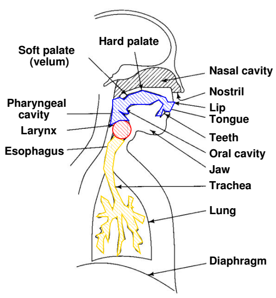
Feature Extraction and Speech Production

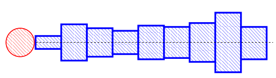


Notes

Source/Filter Model, General Case

Vowels



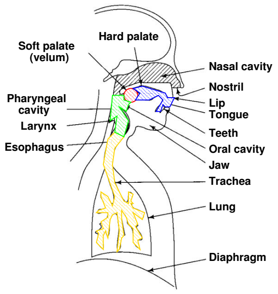


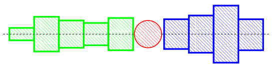
- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Notes

Source/Filter Model, General Case

Fricatives (e.g. sh) or Plosive (e.g. k)



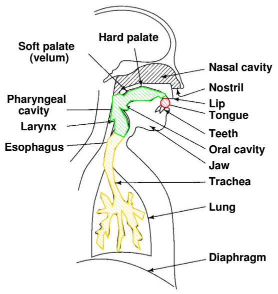


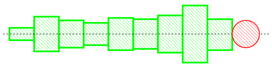
- ☐ Source (noise or impulsive)
- ☐ Front Cavity
- ☐ Back Cavity
- ☐ Back Cavity (2nd approx.)

Notes

Source/Filter Model, General Case

Fricatives (e.g. s) or Plosive (e.g. t)



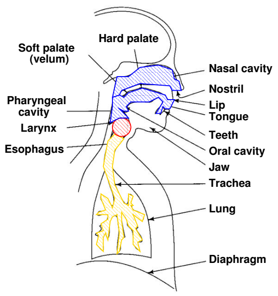


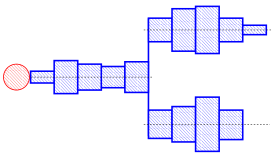
- ☐ Source (noise or impulsive)
- ☐ Front Cavity
- ☐ Back Cavity
- ☐ Back Cavity (2nd approx.)

Notes

Source/Filter Model, General Case

Nasalised Vowels

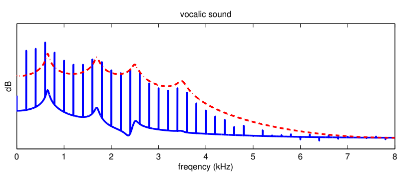


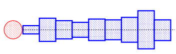


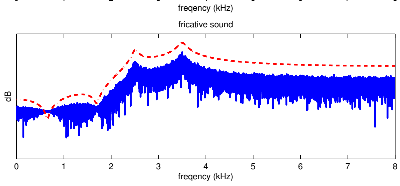
- ☐ Source (periodic)
- ☐ Front Cavity
- ☐ Back Cavity
- ☐ Back Cavity (2nd approx.)

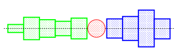
Notes

Examples





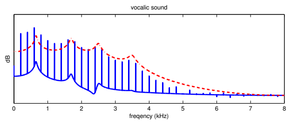




Notes

Relevant vs Irrelevant Information

For the purpose of transcribing words:
Relevant: vocal tract shape → spectral envelope
Irrelevant: vocal fold vibration frequency (f0) → spectral details



Exceptions:

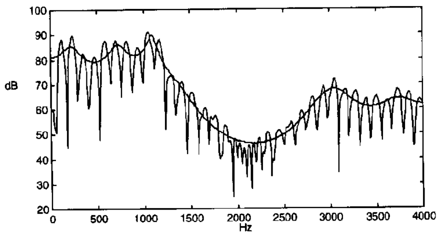
- tonal languages (Chinese)
- pitch and prosody convey meaning

Notes

Linear Prediction Analysis

Attempt to model the vocal tract filter

$$\tilde{x}[n] = \sum_{k=1}^{p} a_k x[n-k]$$



better match at spectral peaks than valleys

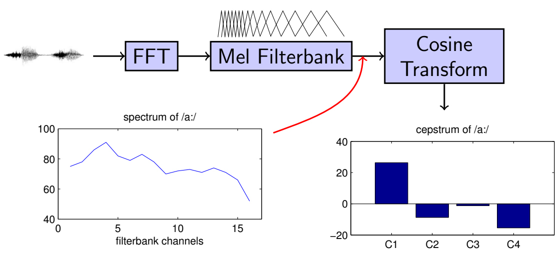
Notes

Mel Frequency Cepstrum Coefficients

- imitate aspects of auditory processing
- *de facto* standard in ASR
- does not assume all-pole model of the spectrum
- uncorrelated: easier to model statistically

Notes

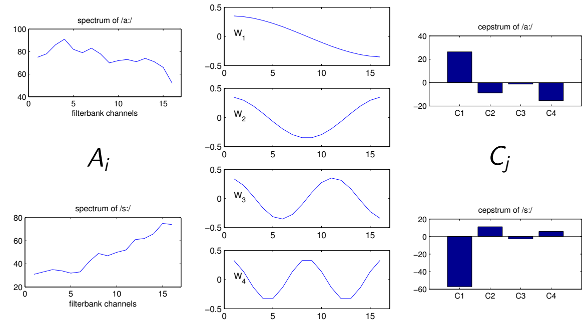
MFCCs Calculation



Notes

Cosine Transform

$$C_j = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} A_i \cos\left(\frac{j\pi(i-0.5)}{N}\right)$$



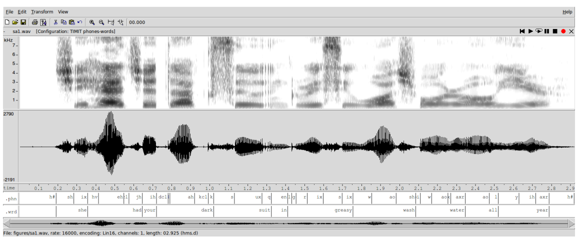
Notes

MFCCs: typical values

- 12 Coefficients C1–C12
- Energy (could be C0)
- Delta coefficients (derivatives in time)
- Delta-delta (second order derivatives)
- total: 39 coefficients per frame (analysis window)

Notes

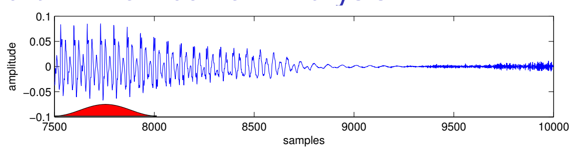
A time varying signal

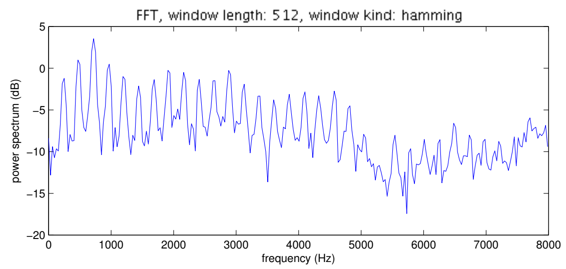


- speech is time varying
- short segment are quasi-stationary
- use short time analysis

Notes

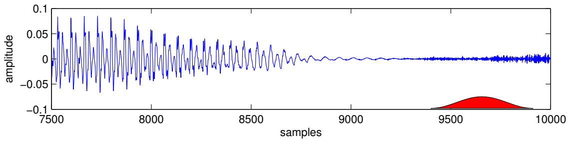
Short-Time Fourier Analysis

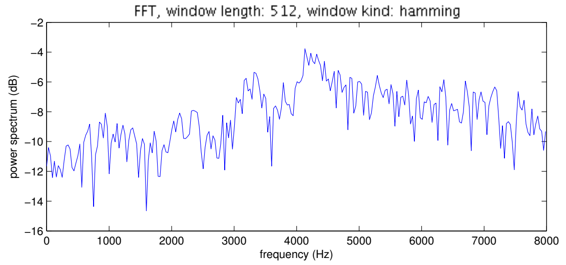




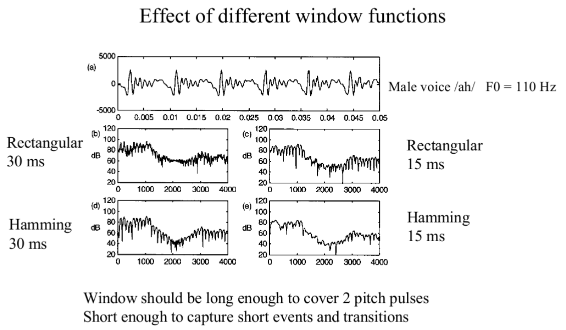
Notes

Short-Time Fourier Analysis





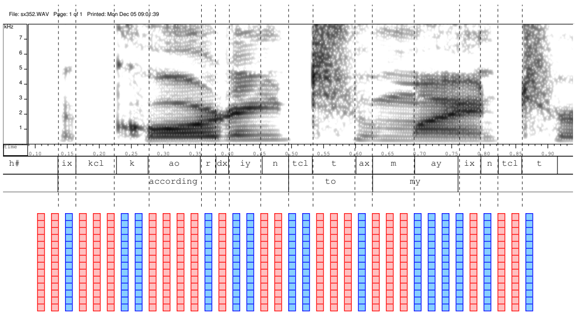
Short-Time Fourier Analysis



Windowing, typical values

- signal sampling frequency: 8–20kHz
- analysis window: 10–50ms
- frame interval: 10–25ms (100–40Hz)

Frame-Based Processing



Notes

Notes

Notes

Notes

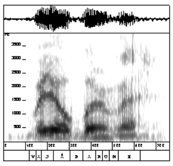
Comparing frames

- city block distance: $d(x, y) = \sum_i |x_i - y_i|$
- Euclidean distance: $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
- Mahalanobis distance:
  $d(x, y) = \sum_i (x_i - \mu_y)^2 / \sigma_y$
- probability function:
  $f(X = x | \mu, \Sigma) = N(x; \mu, \Sigma)$
- artificial neural networks: $d = f\left(\sum_i w_i x_i - \theta\right)$

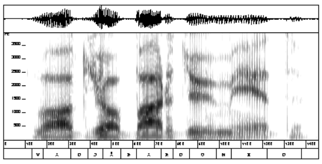
Notes

Comparing Utterances

In order to recognise speech we have to be able to compare different utterances



Va jobbaru me



Vad jobbar du med

Notes

Fixed vs Variable Length Representation

Notes

Combining frame-wise scores into utterance scores

Template Matching

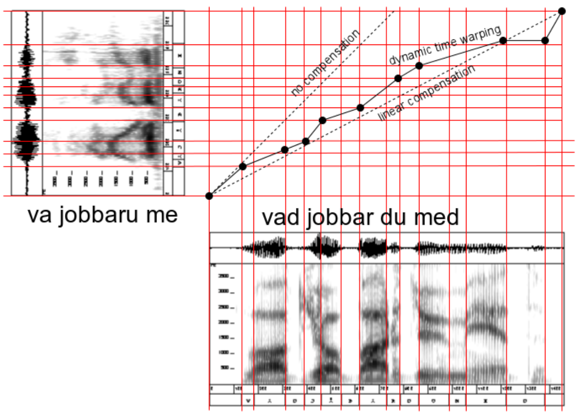
- oldest technique
- simple comparison of template patterns
- compensate for varying speech rate (Dynamic Programming)

Hidden Markov Models (HMMs)

- most used technique
- models of segmental structure of speech
- recognition by Viterbi search (Dynamic Programming)

Notes

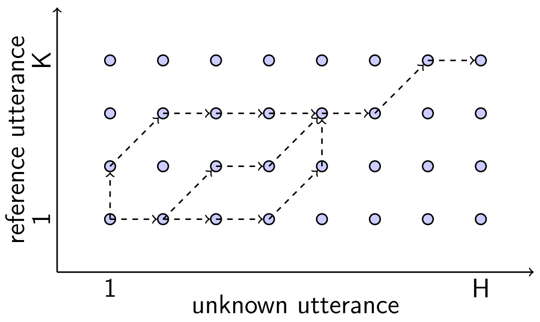
Template Matching



Notes

Dynamic Programming

- compare any possible alignment
- problem: exponential with H and K!

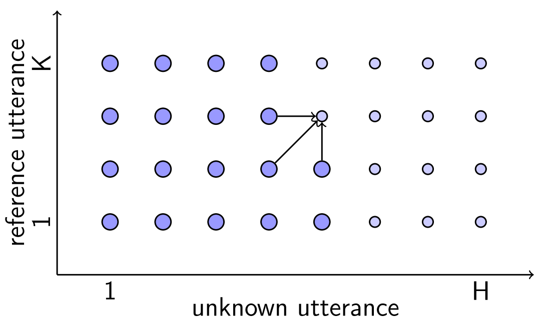


Notes

Dynamic Programming

Dynamic Time Warping (DTW) algorithm

for $h = 1$ to H **do**
  for $k = 1$ to K **do**
    $AccD[h, k] = LocD[h, k] + \min(AccD[h-1, k], AccD[h-1, k-1], AccD[h, k-1])$



Notes

DP Example: Spelling

- observations are letters
- local distance: 0 (same letter), 1 (different letter)
- Unknown utterance: ALLDRIG
- Reference1: ALDRIG
- Reference2: ALLTID
- Problem: find closest match

Distance char-by-char:

- ALLDRIG–ALDRIG = 5
- ALLDRIG–ALLTID = 4

Notes

DP Example: Solution

LocD[h,k]=

G	1	1	1	1	1	1	0
I	1	1	1	1	1	0	1
R	1	1	1	1	0	1	1
D	1	1	1	0	1	1	1
L	1	0	0	1	1	1	1
A	0	1	1	1	1	1	1
	A	L	L	D	R	I	G

AccD[h,k]=

G	5	4	4	3	2	1	0
I	4	3	3	2	1	0	1
R	3	2	2	1	0	1	2
D	2	1	1	0	1	2	3
L	1	0	0	1	2	3	4
A	0	1	2	3	4	5	6
	A	L	L	D	R	I	G

Distance ALLDRIG–ALDRIG: *AccD*[H,K] = 0
Distance ALLDRIG–ALLTID?

Notes

DP Example: Solution

LocD[h,k]=

D	1	1	1	0	1	1	1
I	1	1	1	1	1	0	1
T	1	1	1	1	1	1	1
L	1	0	0	1	1	1	1
L	1	0	0	1	1	1	1
A	0	1	1	1	1	1	1
	A	L	L	D	R	I	G

AccD[h,k]=

D	5	3	3	2	3	3	3
I	4	2	2	2	2	2	3
T	3	1	1	1	2	3	4
L	2	0	0	1	2	3	4
L	1	0	0	1	2	3	4
A	0	1	2	3	4	5	6
	A	L	L	D	R	I	G

Distance ALLDRIG–ALDRIG: *AccD*[H,K] = 0
Distance ALLDRIG–ALLTID: *AccD*[H,K] = 3

Notes

Best path: Backtracking

Sometimes we want to know the path

1. at each point [h,k] remember the minimum distance predecessor (back pointer)
2. at the end point [H,K] follow the back pointers until the start

Notes

Properties of Template Matching

Pros:

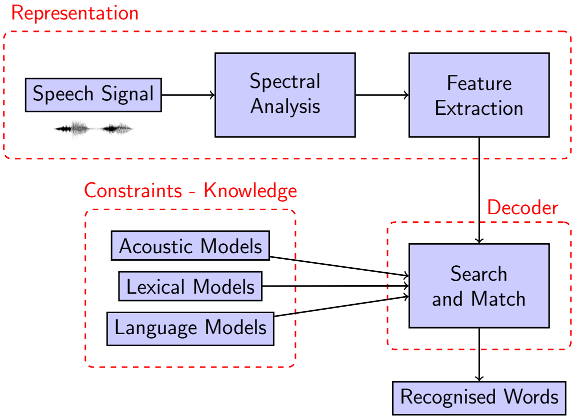
+ No need for phonetic transcriptions
+ within-word co-articulation for free
+ high time resolution

Cons:

- cross-word co-articulation not modelled
- requires recordings of every word
- not easy to model variation
- does not scale up with vocabulary size

Notes

Components of ASR System



Notes

A probabilistic perspective

1. Compute probability of a word sequence given the acoustic observation: $P(\text{words}|\text{sounds})$
2. find the optimal word sequence by maximising the probability:

$$\widehat{\text{words}} = \arg\max P(\text{words}|\text{sounds})$$

Notes

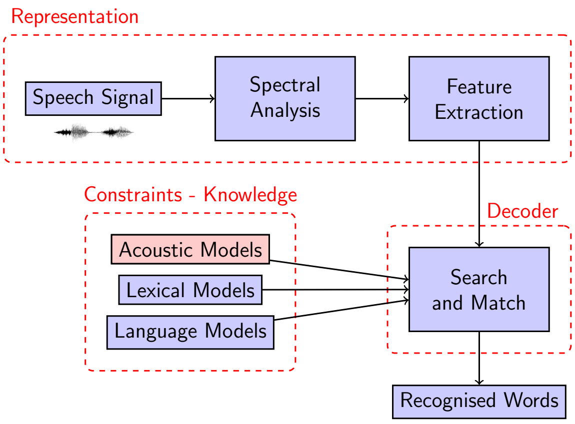
A probabilistic perspective: Bayes' rule

$$P(\text{words}|\text{sounds}) = \frac{P(\text{sounds}|\text{words})P(\text{words})}{P(\text{sounds})}$$

- $P(\text{sounds}|\text{words})$ can be estimated from training data and transcriptions
- $P(\text{words})$: *a priori* probability of the words (Language Model)
- $P(\text{sounds})$: *a priori* probability of the sounds (constant, can be ignored)

Notes

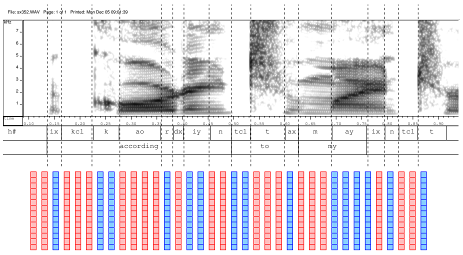
Components of ASR System



Notes

Probabilistic Modelling

Problem: How do we model $P(\text{sounds}|\text{words})$?



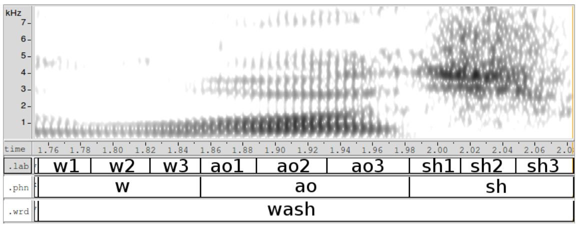
Every feature vector (observation at time t) is a continuous stochastic variable (e.g. MFCC)

Notes

Stationarity

Problem: speech is not stationary

- we need to model short segments independently
- the fundamental unit can not be the word, but must be shorter
- usually we model three segments for each phoneme



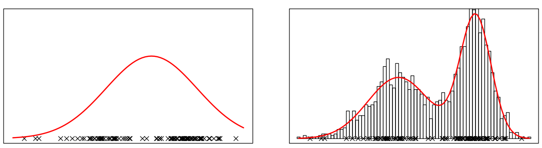
Notes

Local probabilities (frame-wise)

If segment sufficiently short

$$P(\text{sounds}|\text{segment})$$

can be modelled with standard probability distributions
Usually Gaussian or Gaussian Mixture

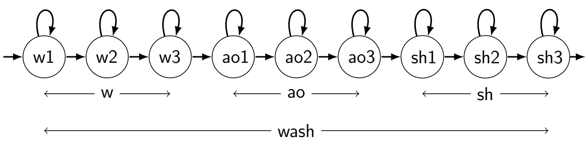


Notes

Global Probabilities (utterance)

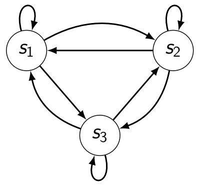
Problem: How do we combine the different $P(\text{sounds}|\text{segment})$ to form $P(\text{sounds}|\text{words})$?

Answer: Hidden Markov Model (HMM)



Notes

Hidden Markov Models (HMMs)

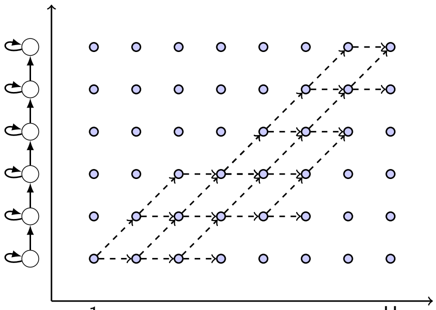


Elements:

set of states:	$S = \{s_1, s_2, s_3\}$
transition probabilities:	$T(s_a, s_b) = P(s_b, t \mid s_a, t-1)$
prior probabilities:	$\pi(s_a) = P(s_a, t_0)$
state to observation probabilities:	$B(o, s_a) = P(o \mid s_a)$

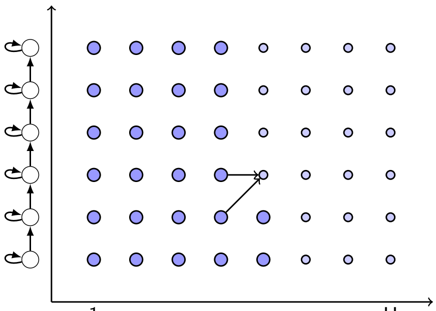
Notes

Hidden Markov Models (HMMs)



Notes

Hidden Markov Models (HMMs)



Notes

HMM-questions

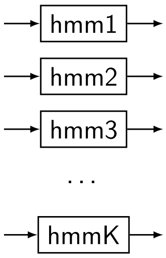
1. what is the probability that the model has generated the sequence of observations? (isolated word recognition) forward algorithm
2. what is the most likely state sequence given the observation sequence? (continuous speech recognition) Viterbi algorithm [5]
3. how can the model parameters be estimated from examples? (training) Baum-Welch[1]

[5] A. J. Viterbi. "Error Bounds for Convolutional Codes and an Asymtotically optimum decoding algorithm". In: *IEEE Trans. Inform. Theory* IT-13 (Apr. 1967), pp. 260–269

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *Ann. Math. Statist.* 41.1 (1970), pp. 164–171

Notes

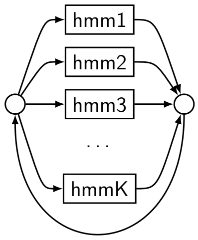
Isolated Words Recognition



Compare Likelihoods (forward-backward)

Notes

Continuous Speech Recognition

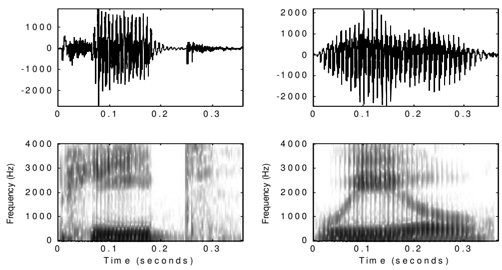


Viterbi algorithm

Notes

Modelling Coarticulation

Example peat /pi:t/ vs wheel /wi:l/



Notes

Modelling Coarticulation

Context dependent models (CD-HMMs)

- Duplicate each phoneme model depending on left and right context:
- from "a" monophone model
- to "d−a+f", "d−a+g", "l−a+s"... triphone models
- If there are $N = 50$ phonemes in the language, there are $N^3 = 125000$ potential triphones
- many of them are not exploited by the language

Notes

Amount of parameters

Example:

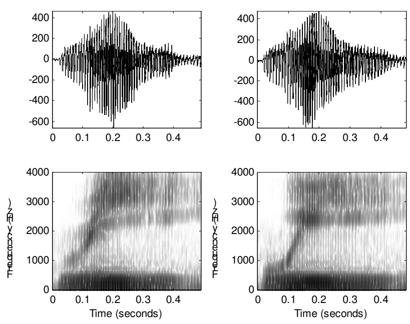
- a large vocabulary recogniser may have 60000 triphone models
- each model has 3 states
- each state may have 32 mixture components with $1 + 39 \times 2$ parameters each (weight, means, variances): $39 \times 32 \times 2 + 32 = 2528$

Totally it is $60000 \times 3 \times 2528 = 455$ million parameters!

Notes

Similar Coarticulation

/riː/ vs /wiː/



Notes

Tying to reduce complexity

Example: similar triphones d−a+m and t−a+m

- same right context, similar left context
- 3rd state is expected to be very similar
- 2nd state may also be similar

States (and their parameters) can be shared between models

\+ reduce complexity

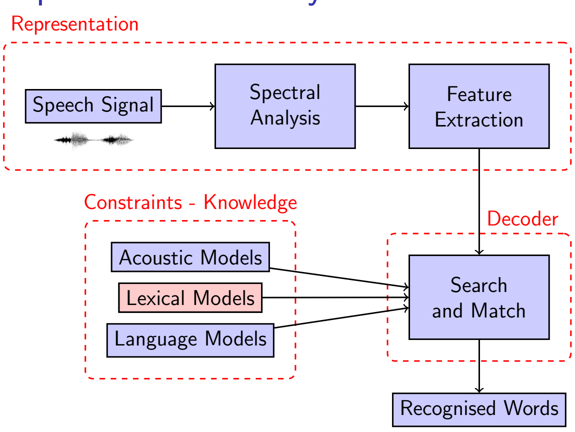
\+ more data to estimate each parameter

− fine detail may be lost

done with CART tree methodology

Notes

Components of ASR System



Notes

Lexical Models

- in general specify sequence of phoneme for each word
- example:

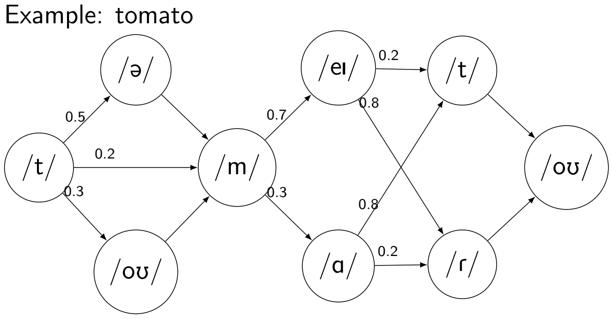
	IPA	X-SAMPA
"dictionary"		
UK:	/d ɪ k ʃ ə n (ə) ɹ i/	/d I k S @ n (@) r i/
USA:	/d ɪ k ʃ ə n ɛ ɹ i/	/d I k S @ n E r i/

- expensive resources
- include multiple pronunciations
- phonological rules (assimilation, deletion)

Notes

Pronunciation Network

Example: tomato



Notes

Assimilation

did you	/d ɪ dʒ j ə/
set you	/s ɛ tʃ ɜ/
last year	/l æ s tʃ iː ɹ/
because you've	/b iː k ə ʒ uː v/

Notes

Deletion

find him	/f a ɪ n ɪ m/
around this	/ə ɹ aʊ n ɪ s/
let me in	/l ɛ m iː n/

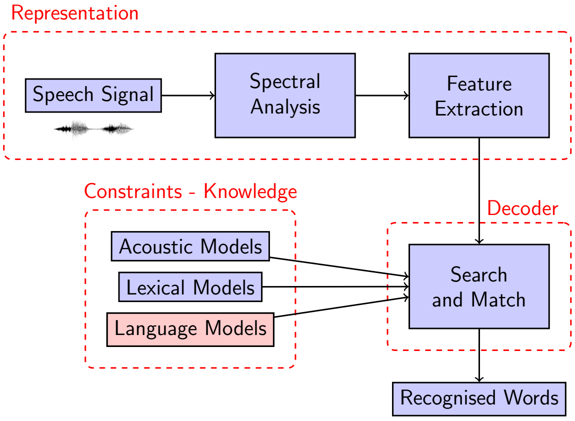
Notes

Out of Vocabulary Words

- Proper names often not in lexicon
- derive pronunciation automatically
- English has very complex grapheme-to-phoneme rules
- attempts to derive pronunciation from speech recordings

Notes

Components of ASR System



Notes

Why do we need language models?

Bayes' rule:

$$P(\text{words}|\text{sounds}) = \frac{P(\text{sounds}|\text{words})P(\text{words})}{P(\text{sounds})}$$

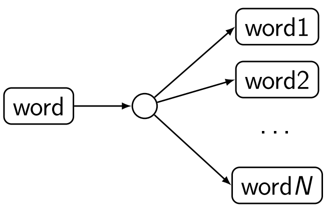
where
$P(\text{words})$: *a priori* probability of the words (Language Model)

We could use non informative priors ($P(\text{words}) = 1/N$), but...

Notes

Branching Factor

- if we have N words in the dictionary
- at every word boundary we have to consider N equally likely alternatives
- N can be in the order of millions



Notes

Ambiguity

"ice cream" vs "I scream"
/aɪ s k ɹ iː m/

Notes

Language Models

$$P(\text{words}|\text{sounds}) = \frac{P(\text{sounds}|\text{words})P(\text{words})}{P(\text{sounds})}$$

Finite state networks (hand-made, see lab)

- formal language, e.g. traffic control

Statistical Models (N-grams)

- unigrams: $P(w_i)$
- bigrams: $P(w_i|w_{i-1})$
- trigrams: $P(w_i|w_{i-1}, w_{i-2})$
- ...

Notes

Chomsky's formal grammar

Noam Chomsky: linguist, philosopher, ...

$$G = (V, T, P, S)$$

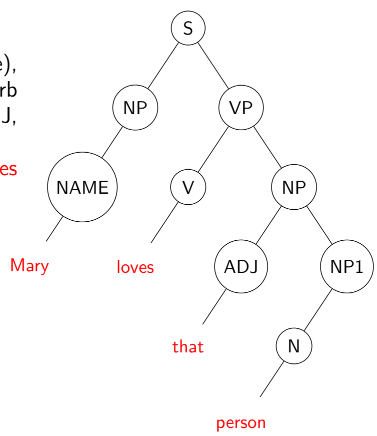
where

V: set of non-terminal constituents
T: set of terminals (lexical items)
P: set of production rules
S: start symbol

Notes

Example

$S =$ sentence
$V =$ {NP (noun phrase), NP1, VP (verb phrase), NAME, ADJ, V (verb), N (noun)}
$T =$ {Mary , person , loves , that , ...}
$P =$ {S → NP VP
NP → NAME
NP → ADJ NP1
NP1 → N
VP → VERB NP
NAME → Mary
V → loves
N → person
ADJ → that }



Notes

Formal Language Models

- only used for simple tasks
- hard to code by hand
- people do not speak following formal grammars

Notes

Statistical Grammar Models (N-grams)

Simply count co-occurrence of words in large text data sets

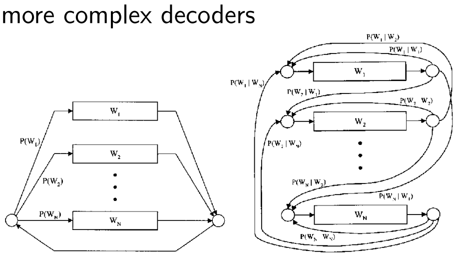
- unigrams: $P(w_i)$
- bigrams: $P(w_i|w_{i-1})$
- trigrams: $P(w_i|w_{i-1}, w_{i-2})$
- ...

Notes

Language Models: complexity

Increasing N in N-grams leads to:

1. more complex decoders



2. difficulties in training the LM parameters

Notes

Knowledge Models in ASR

Acoustic Models trained on hours of annotated speech recordings (especially developed speech databases)

Lexical Model usually produced by hand by experts (or generated by rules)

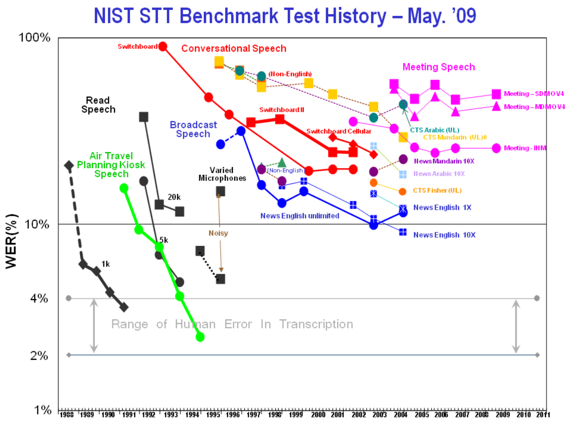
Language Models trained on millions of words of text (often from news papers)

Notes

Main variables in ASR

- **Speaking mode** isolated words vs continuous speech
- **Speaking style** read speech vs spontaneous speech
- **Speakers** speaker dependent vs speaker independent
- **Vocabulary** small (<20 words) vs large (>50 000 words)
- **Robustness** against background noise

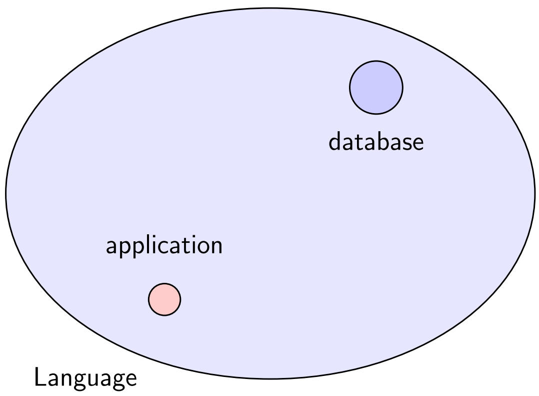
Notes



http://www.itl.nist.gov/iad/mig/publications/ASRhistory/

Notes

Why is it so hard?



Notes

Challenges — Variability

Between speakers
- Age
- Gender
- Anatomy
- Dialect

Within speaker
- Stress
- Emotion
- Health condition
- Read vs Spontaneous
- Adaptation to environment (Lombard effect)
- Adaptation to listener

Environment
- Noise
- Room acoustics
- Microphone distance
- Microphone, telephone
- Bandwidth

Listener
- Age
- Mother tongue
- Hearing loss
- Known / unknown
- Human / Machine

Notes

Sheep and Goats [3]

[3] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. "SHEEP, GOATS, LAMBS and WOLVES A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation". In: *INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING*. 1998

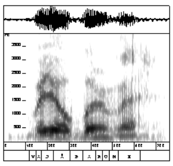
Notes

Sheep and Goats [3]

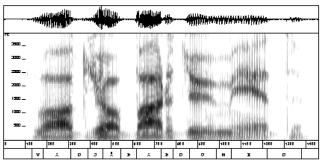
[3] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. "SHEEP, GOATS, LAMBS and WOLVES A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation". In: *INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING*. 1998

Notes

Exmpl: spontaneous vs hyper-articulated



Va jobbaru me



Vad jobbar du med

"What is your occupation"
("What work you with")

Notes

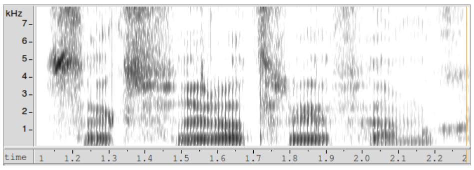
Examples of reduced pronunciation

Spoken	Written	In English
Tesempel	Till exempel	for example
åhamba	och han bara	and he just
bafatt	bara för att	just because
javende	jag vet inte	I don't know

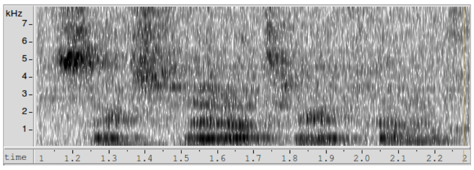
Notes

Microphone distance

Headset



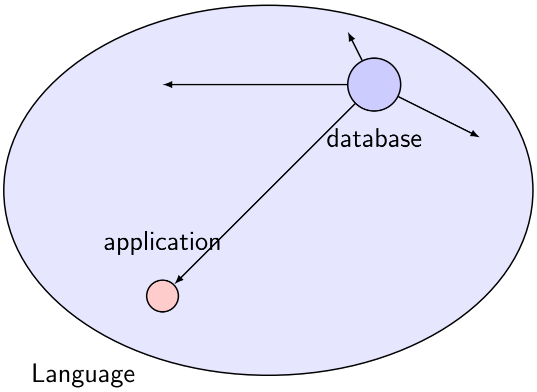
2 m distance



Notes

How do we cope with variability?

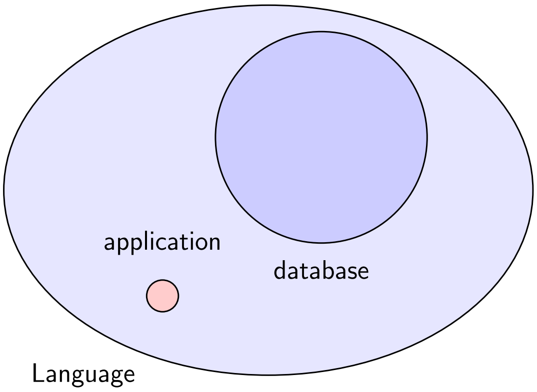
Ideally: models that generalise



Notes

How do we cope with variability?

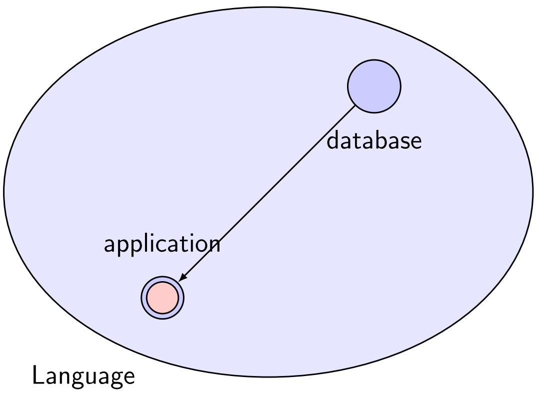
Large companies use insane quantities of data



Notes

How do we cope with variability?

Adaptation



Notes

Adaptation: Example

Enrolment in Dictation Systems

- let the user read a small text before using the system

Beta version of smartphone applications

- the company has all the rights on data generated

Notes

Word Accuracy

$$A = 100\frac{N - S - D - I}{N}$$

Where

- N: total number of reference words
- S: substitutions
- D: deletions
- I: insertions

Notes

Word Accuracy: example

Ref/Rec	I	wanted	badly	to	meet	you
I	corr					
really	del					
wanted		corr				
to			ins	corr		
see					sub	
you						corr

6 words, 1 substitution, 1 insertion, 1 deletion

$$A = 100\frac{6 - 1 - 1 - 1}{6} = 50\%$$

requires dynamic programming

Notes

Measure Difficulty

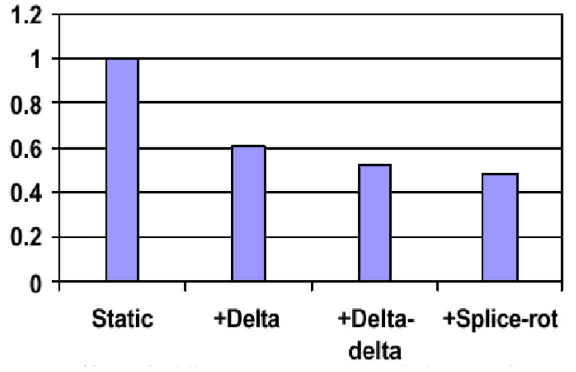
Language Perplexity

$$B = 2^H, \quad H = -\sum_{\forall W} P(W) \log_2(P(W))$$

- $P(W)$ is the probability of the word sequence (language model)
- H is called entropy
- B can be seen as measure of average number of words that can follow any given word
- Example: equiprobable digit sequences $B = 10$

Notes

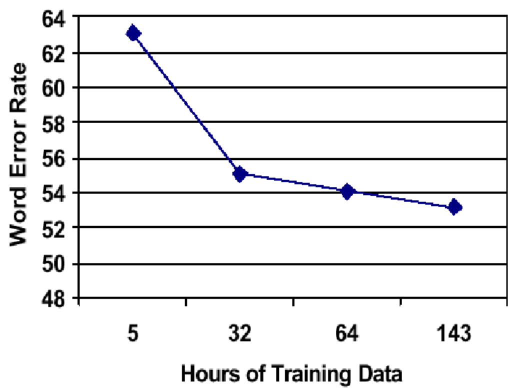
Effect of adding features



Notes

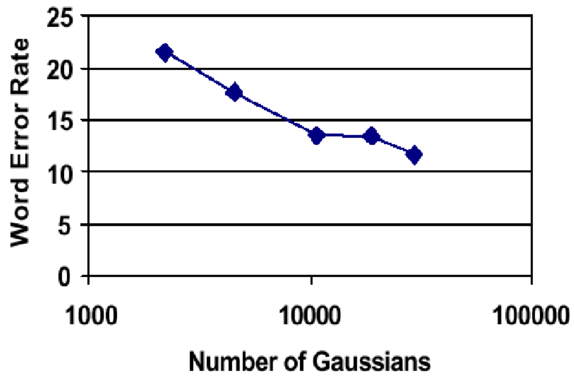
Effect of adding training data

Swichboard data



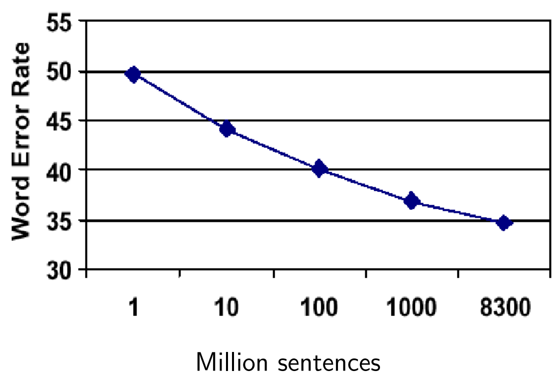
Notes

Effect of adding Gaussians



Notes

Effect of adding data for language models



Notes

Some dictation systems

- vocabulary over 100 000 words
- many languages
- systems: Nuance NatuallySpeaking, Microsoft, (IBM ViaVoice), (Dragon Dictate)

Notes

New applications

- Indexing of TV and radio programs (offline), Google
- real-time subtitling of TV programs (re-speaker that summarises)
- language learning
- smart phones

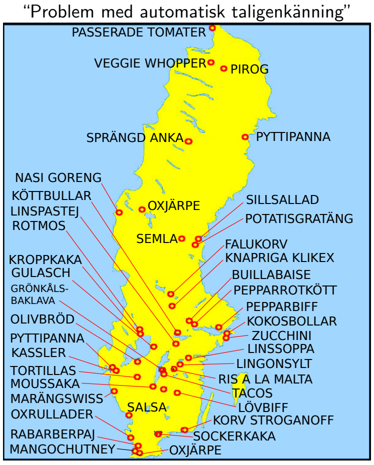
Notes

Limitations

- lack of context
- require huge amounts of training data

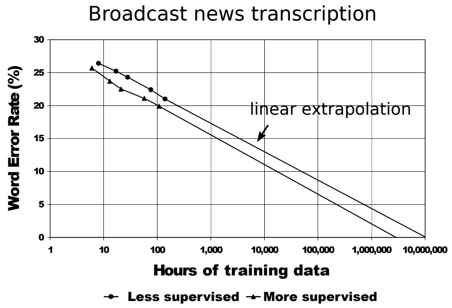
Notes

Adapted from Mikael Parkvall's Lingvistiska Samlarbilder, Nr.96:
"Problem med automatisk taligenkänning"



Notes

Lack of Generalisation[4]



In order to reach 10-years-old's performance,
ASR needs 4 to 70 human lifetimes
exposure to speech!!

[4] R. Moore. "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human Listeners". In: *Proc. of Eurospeech*. Geneva, Switzerland, 2003, pp. 2582–2584

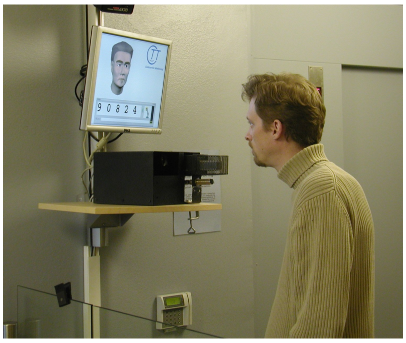
Notes

New directions

- Production inspired modelling
- Study children's speech acquisition
- Modelling and decision techniques
  - Eigenvoices
  - Deep learning neural networks

Notes

Speaker Recognition



Created by Håkan Melin

Notes

Person Identification

Methods rely on:

- something you posses:
  key, magnetic card, . . .
- something you know:
  PIN-code, password, . . .
- something you are:
  physical attributes, behaviour (biometrics)

Notes

Recognition, Verification, Identification

Recognition: general term

Speaker verification:

- an identity is claimed and is verified by voice
- binary decision (accept/reject)
- performance independent of number of users

Speaker identification:

- choose one of N speakers
- close set: voice belongs to one of the N speakers
- open set: any person can access the system
- problem difficulty increases with N

Notes

Text Dependence

Either fix the content or recognise it. Examples:

- Fixed password (text dependent)
- User-specific password
- System prompts the text (prevents impostors from recording and playing back the password)
- any word is allowed (text independent)

text independent

Notes

Representations

Speech Recognition:

- represent speech content
- disregard speaker identity

Speaker Recognition:

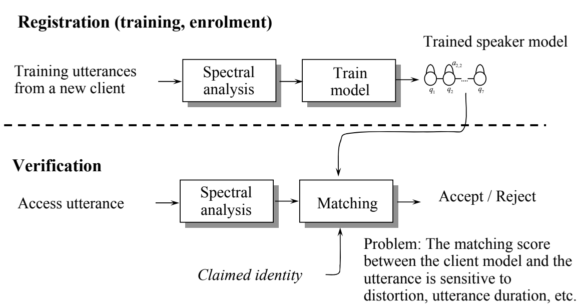
- represent speaker identity
- disregard speech content

Surprisingly:

- MFCCs used for both
- suggests that feature extraction could be improved

Notes

Speaker Verification



Notes

Modelling Techniques

HMMs

- Text dependent systems
- state sequence represents allowed utterance

GMMs (Gaussian Mixture Models)

- Text independent systems
- large number of Gaussian components
- sequential information not used

SVM (Support Vector Machines)

Combined models

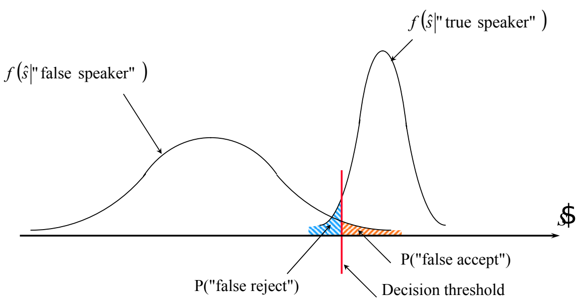
Notes

Evaluation

Claimed Identity	Decision: Accept	Reject
True	OK	False Reject (FR)
False	False Accept (FA)	OK

Notes

Score Distribution and Error Balance



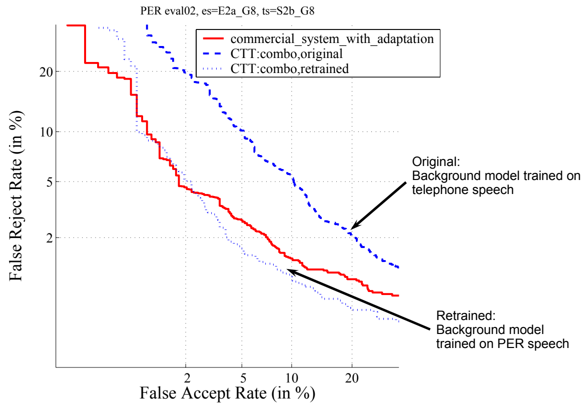
Notes

Performance Measures

- False Rejection Rate (FR)
- False Acceptance Rate (FA)
- Half Total Error Rate (HTER = (FR+FA)/2)
- Equal Error Rate (EER)
- Detection Error Trade-off (DET) Curve

Notes

PER vs Commercial System



Notes

More information and mathematical formulations in DT2118

Notes

Notes

Notes