DT2112
Speech Recognition by Computers

Giampiero Salvi

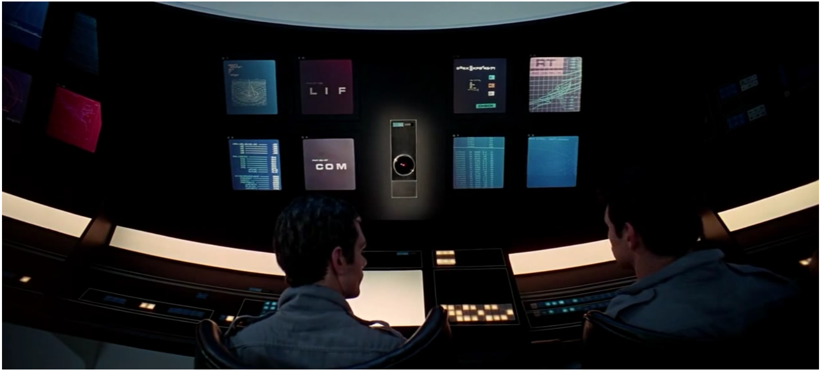
KTH/CSC/TMH giampi@kth.se

VT 2015

Motivation

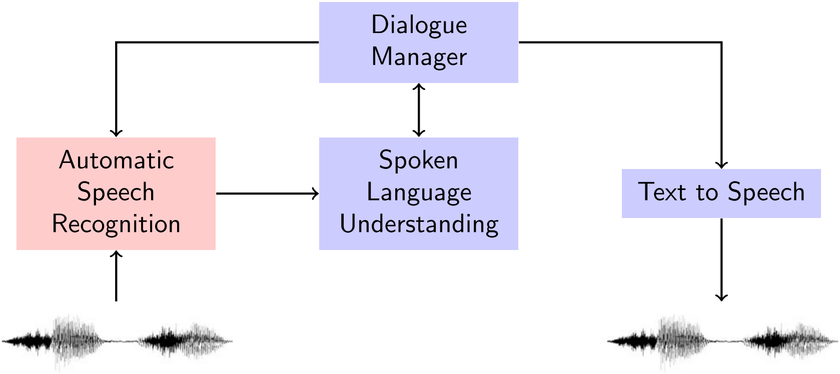
- Natural way of communication (No training needed)
- Leaves hands and eyes free (Good for functionally disabled)
- Effective (Higher data rate than typing)
- Can be transmitted/received inexpensively (phones)

A dream of Artificial Intelligence



2001: A space odyssey (1968)

ASR in a Broader Context



The ASR Scope

Convert speech into text

⟶ Automatic Speech Recognition ⟶ "My name is . . ."
[confidence score]

The ASR Scope

Convert speech into text

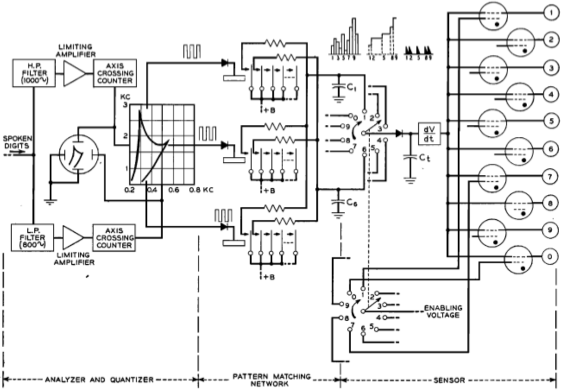
⟶ Automatic Speech Recognition ⟶ "My name is . . ." [confidence score]

Not considered here:

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

A very long endeavour

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

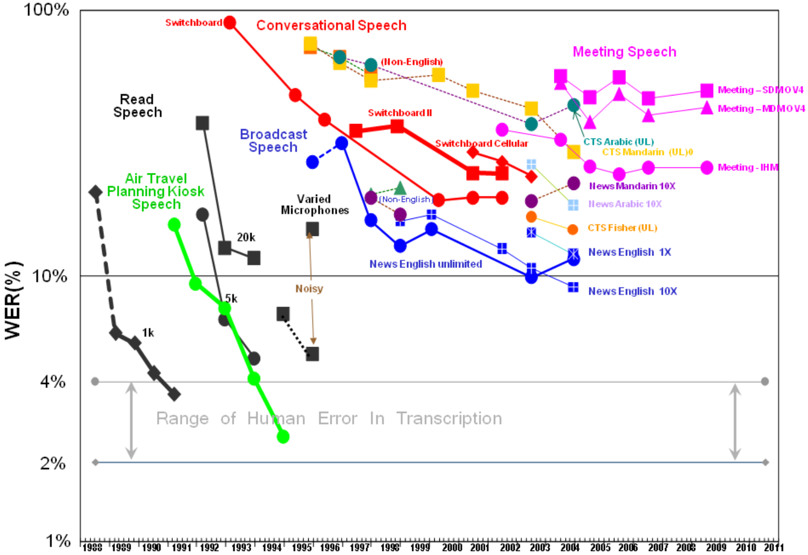


[2] K. H. Davis, R. Biddulph, and S. Balashek. "Automatic Recognition of Spoken Digits". In: *JASA* 24.6 (1952), pp. 637–642

An underestimated challenge

for 60 years many bold
announcements

NIST STT Benchmark Test History – May. '09



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

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

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

Applications today

Call centers:

- traffic information
- time-tables
- booking...

Accessibility

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

Smart phones

- Siri, Android...

Outline

Speech Signal Representations

Template Matching

Probabilistic Approach
- Knowledge Modelling

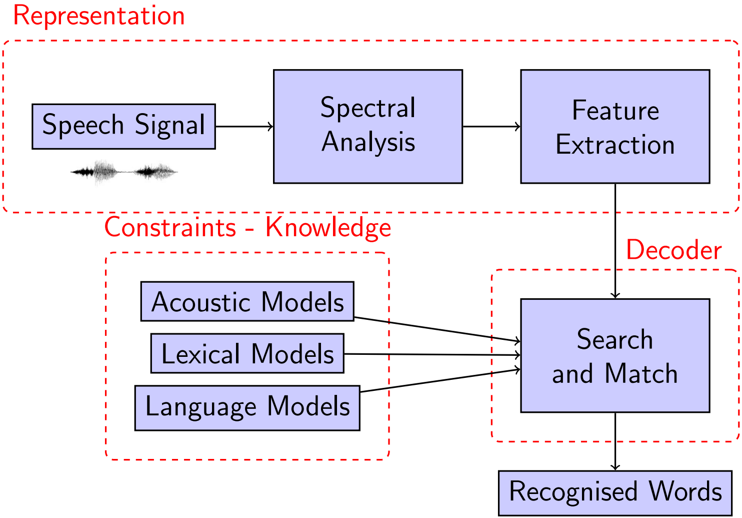
Performance Measures

Robustness and Adaptation

Speaker Recognition

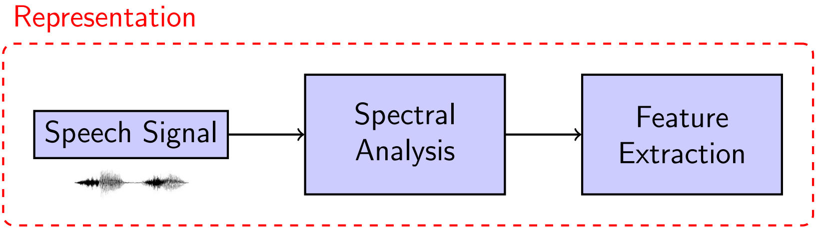
More details in DT2118: “Speech and Speaker Recognition”

Components of ASR System



Outline

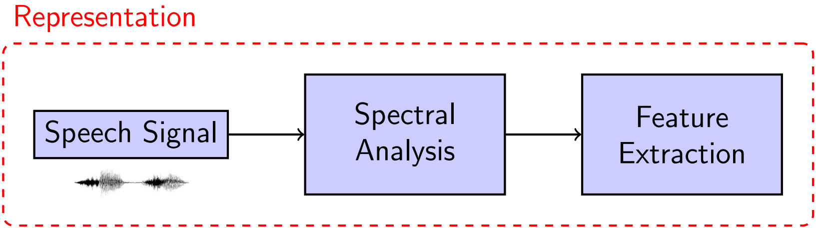
Speech Signal Representations



Goals:

- disregard irrelevant information
- optimise relevant information for modelling

Speech Signal Representations



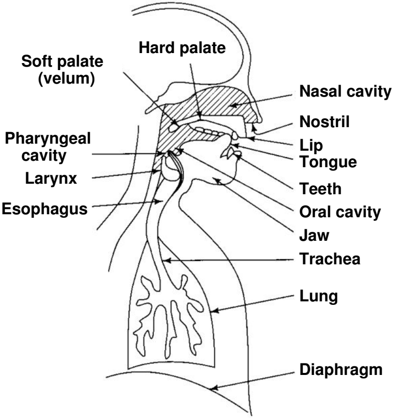
Means:

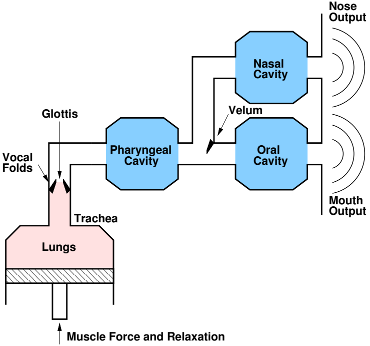
- try to model essential aspects of speech production
- imitate auditory processes
- consider properties of statistical modelling

Examples of Speech Sounds

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

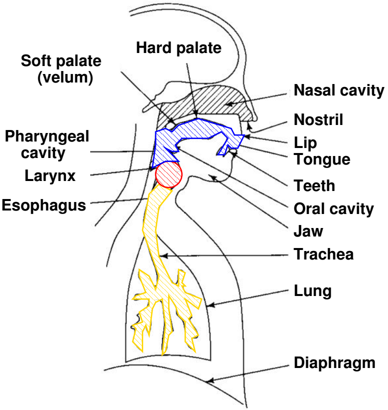
Feature Extraction and Speech Production

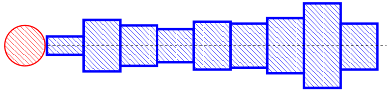




Source/Filter Model, General Case

Vowels

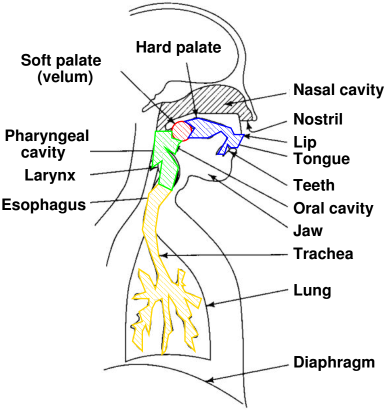


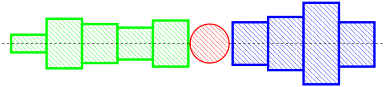


- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

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

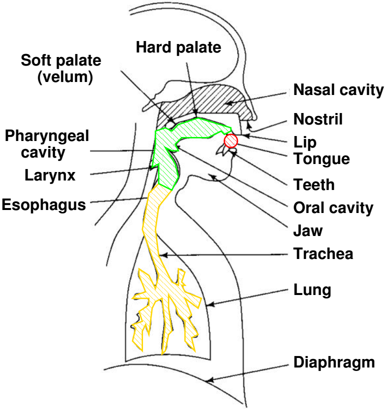


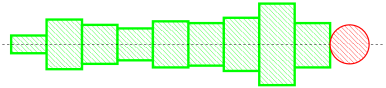


- □ Source (noise or impulsive)
- □ Front Cavity
- □ Back Cavity
- □ Back Cavity (2nd approx.)

Source/Filter Model, General Case

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

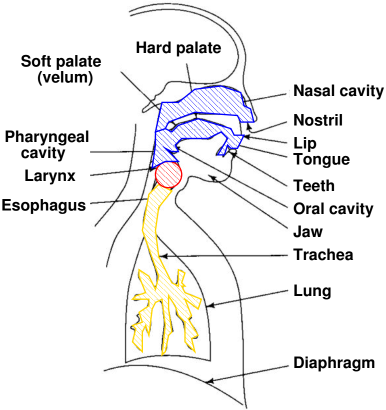


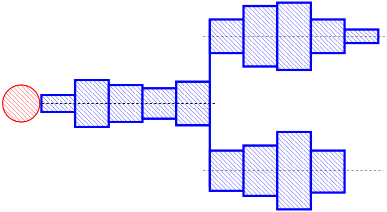


- Source (noise or impulsive)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Source/Filter Model, General Case

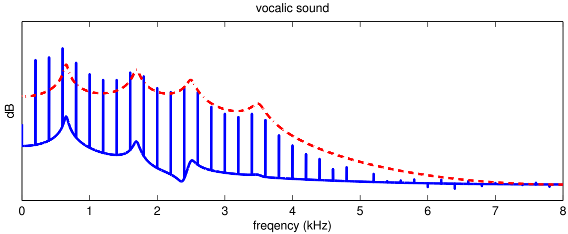
Nasalised Vowels

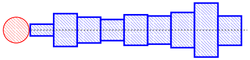


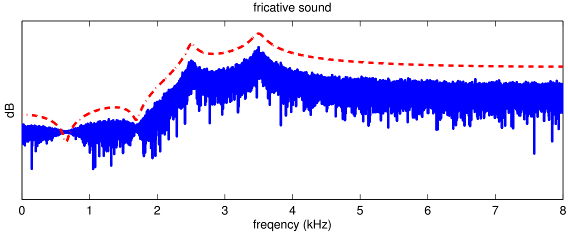


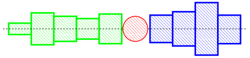
- Source (periodic)
- Front Cavity
- Back Cavity
- Back Cavity (2nd approx.)

Examples


vocalic sound
dB
0
1
2
3
4
5
6
7
8
freqency (kHz)

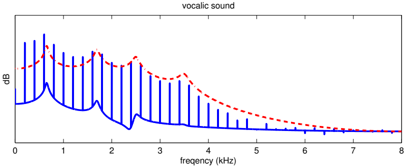



fricative sound
dB
0
1
2
3
4
5
6
7
8
freqency (kHz)



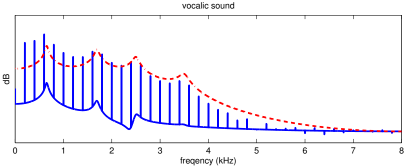
Relevant vs Irrelevant Information

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



Relevant vs Irrelevant Information

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



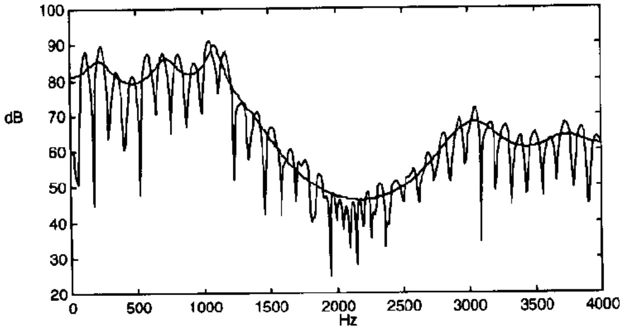
Exceptions:

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

Linear Prediction Analysis

Attempt to model the vocal tract filter

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

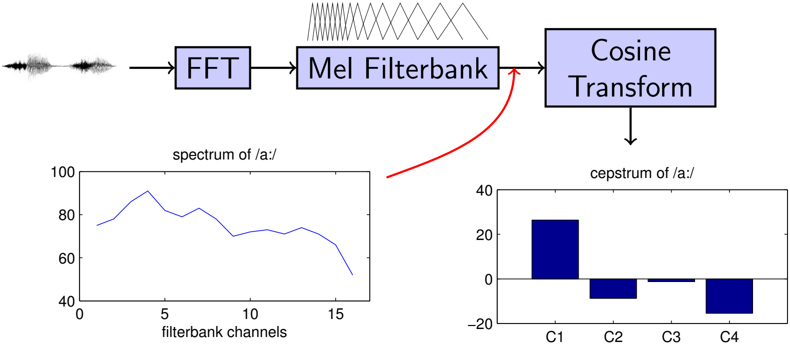


better match at spectral peaks than valleys

Mel Frequency Cepstrum Coefficients

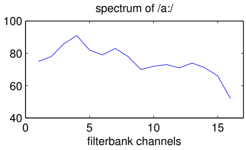
- imitate aspects of auditory processing
- *de facto* standard in ASR
- does not assume all-pole model of the spectrum
- uncorrelated: easier to model statistically

MFCCs Calculation

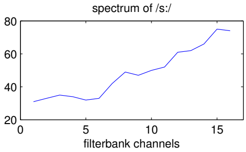


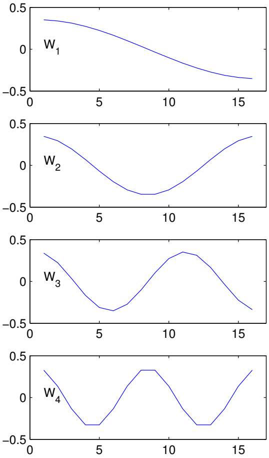
Cosine Transform

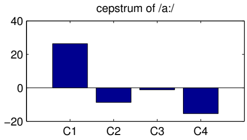
$$C_j = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} A_i \cos\left(\frac{j\pi(i-0.5)}{N}\right)$$



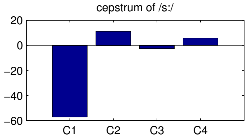
A_i







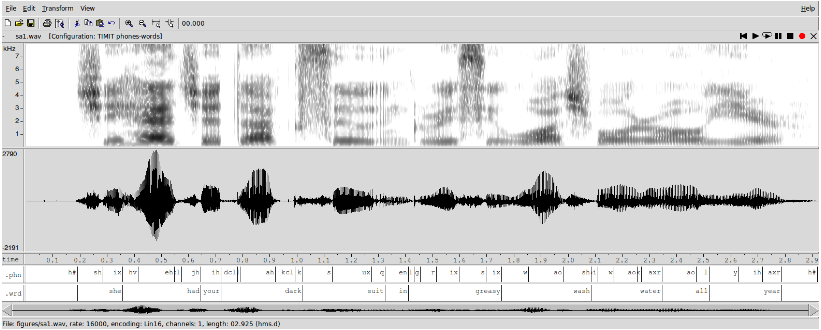
C_j



MFCCs: typical values

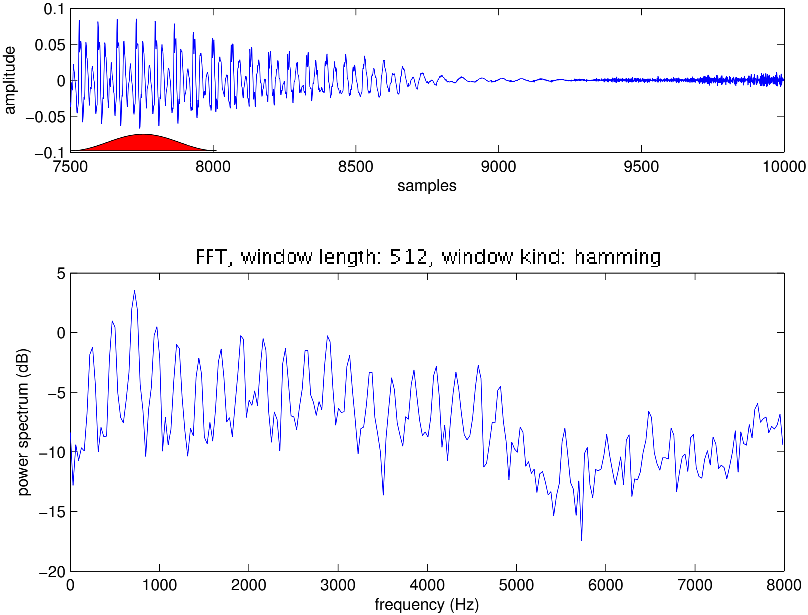
- 12 Coefficients C1–C12
- Energy (could be C0)
- Delta coefficients (derivatives in time)
- Delta-delta (second order derivatives)
- total: 39 coefficients per frame (analysis window)

A time varying signal

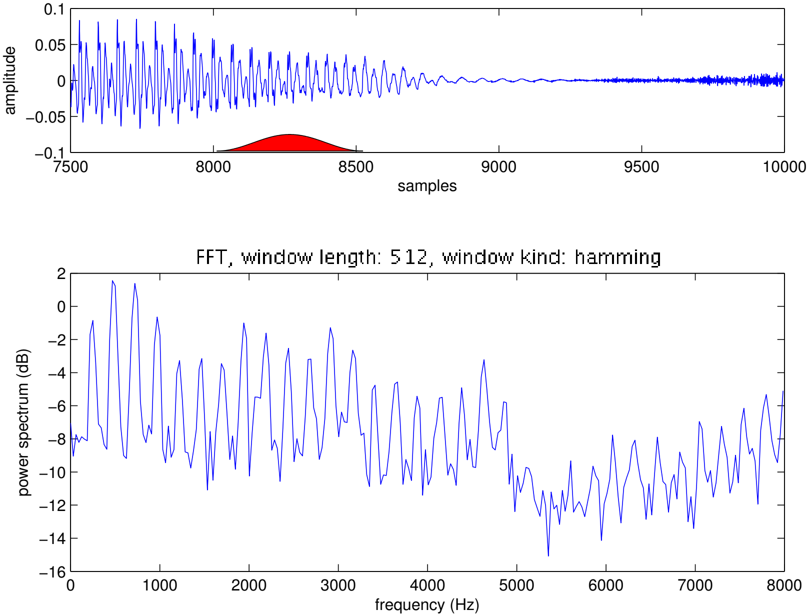


- speech is time varying
- short segment are quasi-stationary
- use short time analysis

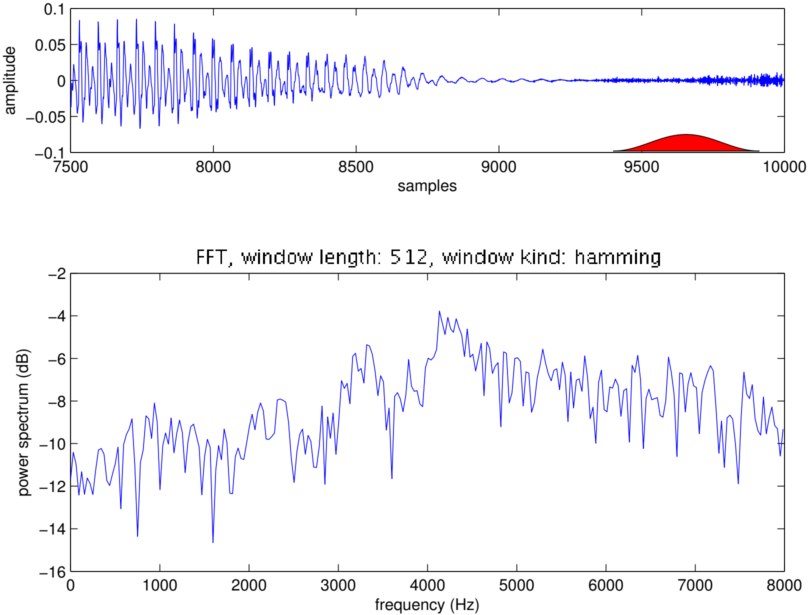
Short-Time Fourier Analysis



Short-Time Fourier Analysis

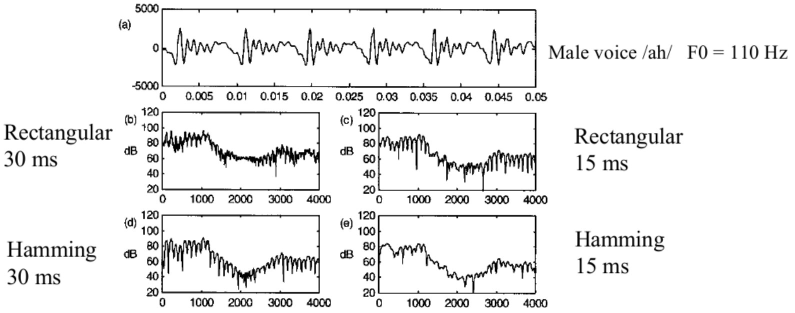


Short-Time Fourier Analysis



Short-Time Fourier Analysis

Effect of different window functions

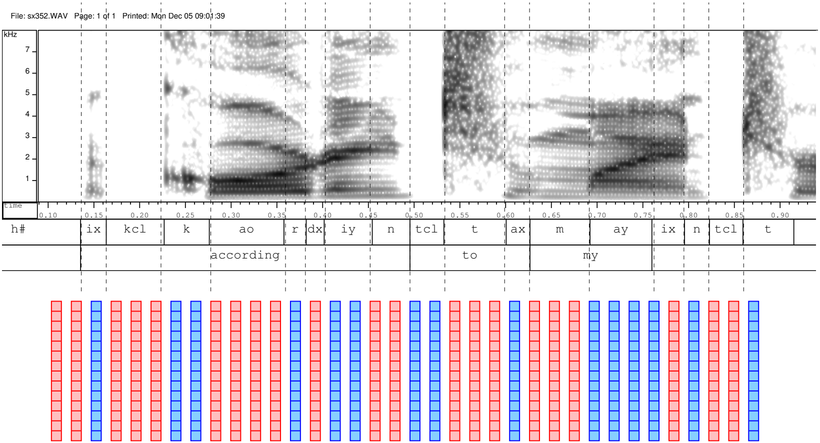


Window should be long enough to cover 2 pitch pulses
Short enough to capture short events and transitions

Windowing, typical values

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

Frame-Based Processing

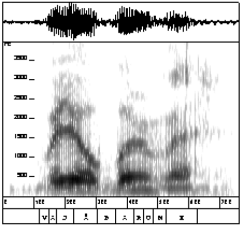


Comparing frames

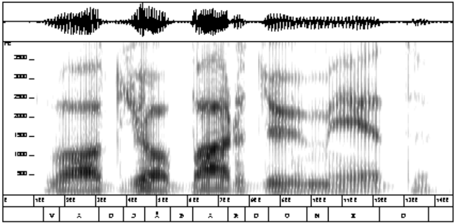
- city block distance: $d(x, y) = \sum_i |x_i - y_i|$
- Euclidean distance: $d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$
- Mahalanobis distance:
  $d(x, y) = \sum_i (x_i - \mu_y)^2 / \sigma_y$
- probability function:
  $f(X = x|\mu, \Sigma) = N(x; \mu, \Sigma)$
- artificial neural networks: $d = f\left(\sum_i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

Comparing Utterances

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



Va jobbaru me



Vad jobbar du med

Fixed vs Variable Length Representation

Combining frame-wise scores into utterance scores

Template Matching

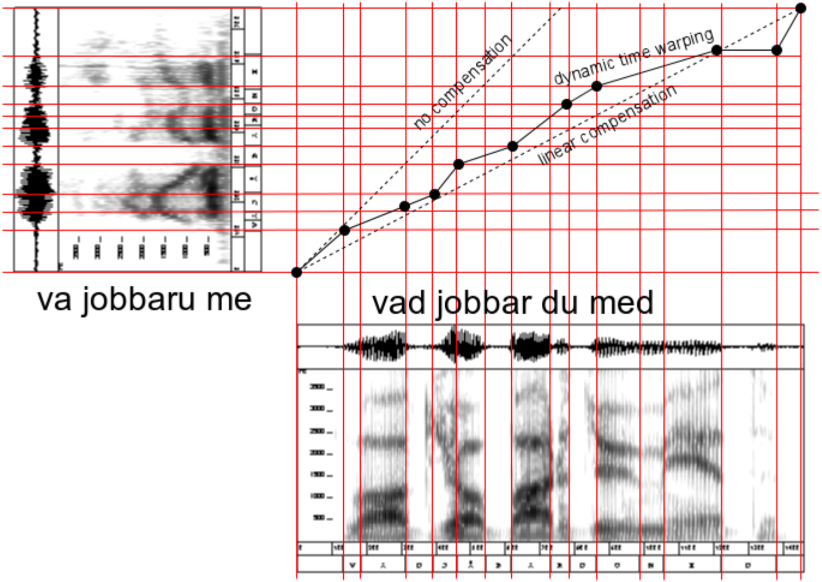
- oldest technique
- simple comparison of template patterns
- compensate for varying speech rate (Dynamic Programming)

Hidden Markov Models (HMMs)

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

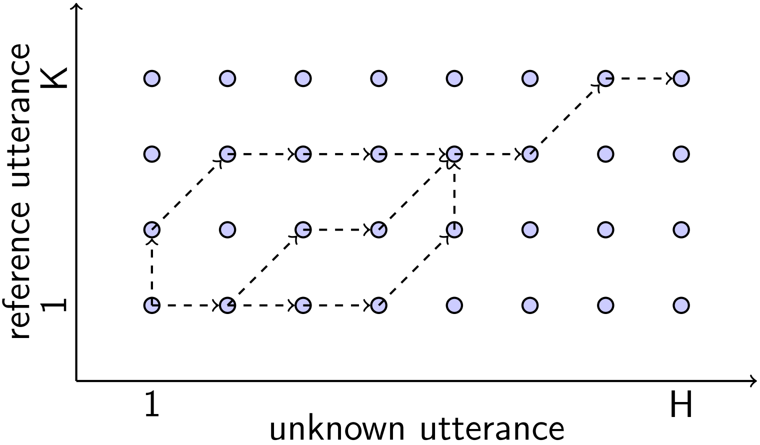
Outline

Template Matching



Dynamic Programming

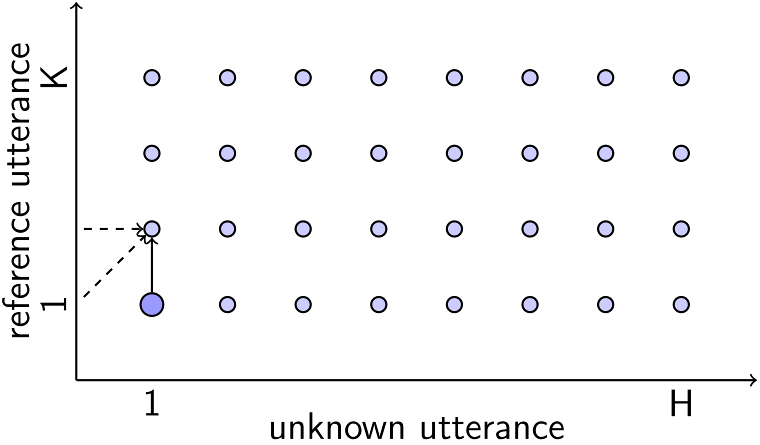
- compare any possible alignment
- problem: exponential with H and K!



Dynamic Programming

Dynamic Time Warping (DTW) algorithm

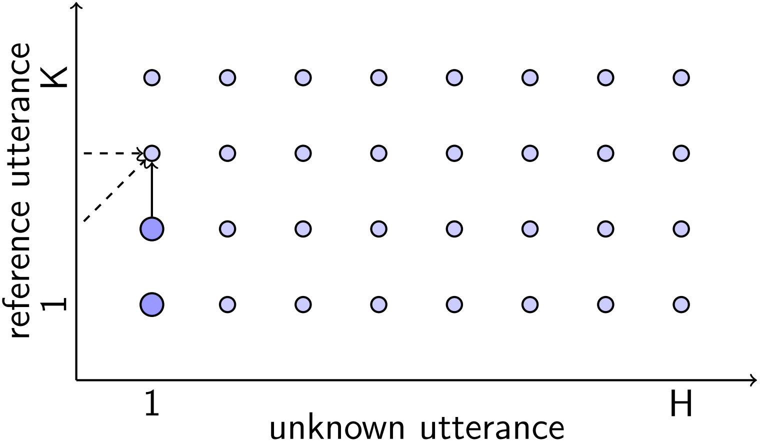
1: **for** $h = 1$ to H **do**
2: **for** $k = 1$ to K **do**
3: $AccD[h,k] = LocD[h,k] + \min(AccD[h-1,k],$
$AccD[h-1,k-1], AccD[h,k-1])$



Dynamic Programming

Dynamic Time Warping (DTW) algorithm

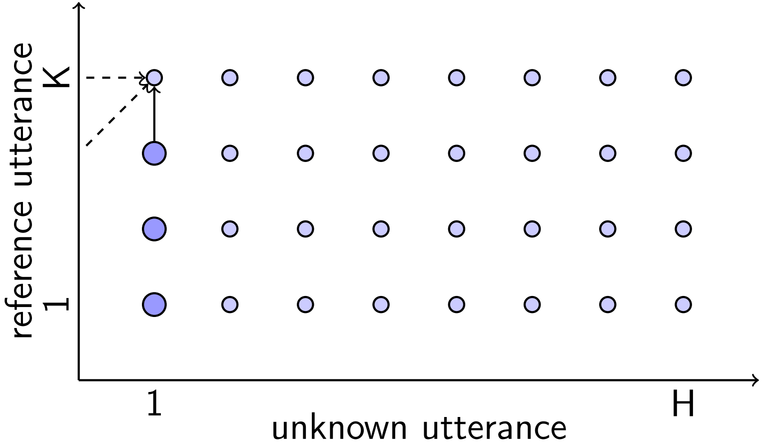
1: **for** $h = 1$ to H **do**
2: **for** $k = 1$ to K **do**
3: $AccD[h, k] = LocD[h, k] + \min(AccD[h-1, k], AccD[h-1, k-1], AccD[h, k-1])$



Dynamic Programming

Dynamic Time Warping (DTW) algorithm

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



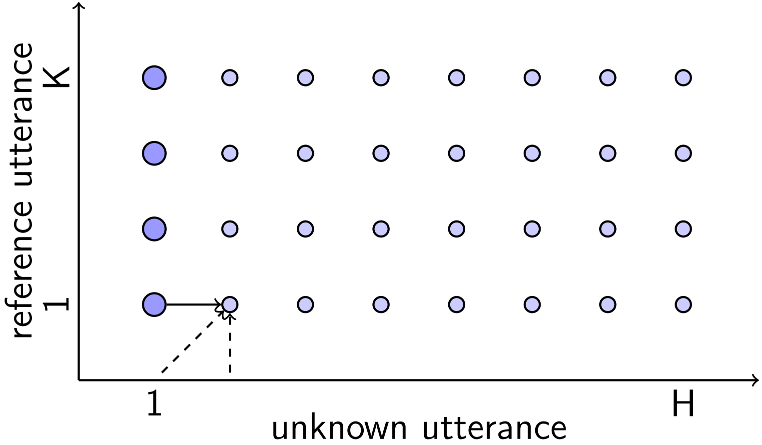
Dynamic Programming

Dynamic Time Warping (DTW) algorithm

1: **for** $h = 1$ to H **do**

2: **for** $k = 1$ to K **do**

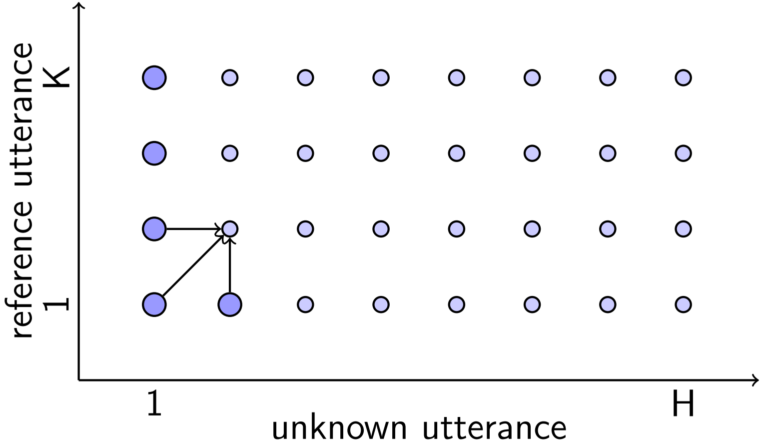
3: $AccD[h,k] = LocD[h,k] + \min(AccD[h-1,k],$
$AccD[h-1,k-1], AccD[h,k-1])$



Dynamic Programming

Dynamic Time Warping (DTW) algorithm

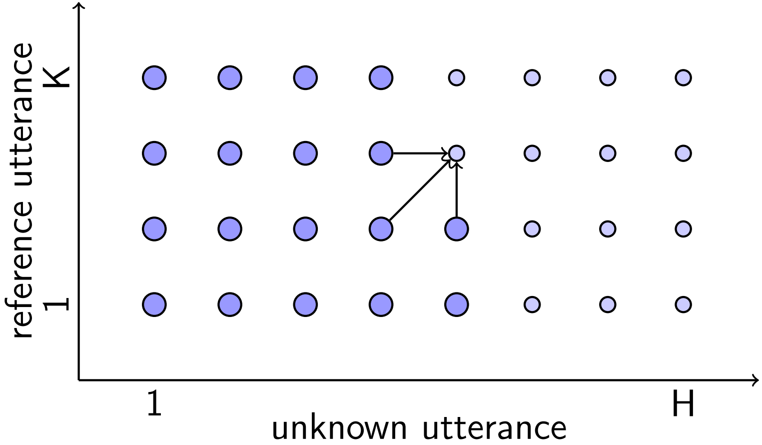
1: **for** $h = 1$ to H **do**
2: **for** $k = 1$ to K **do**
3: $AccD[h,k] = LocD[h,k] + \min(AccD[h-1,k], AccD[h-1,k-1], AccD[h,k-1])$



Dynamic Programming

Dynamic Time Warping (DTW) algorithm

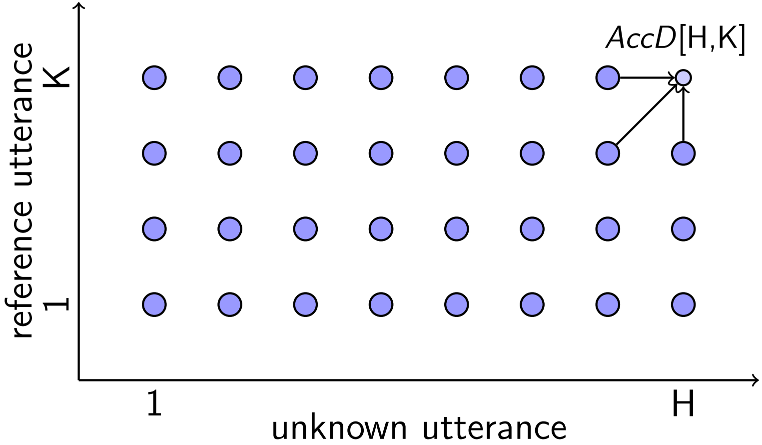
1: **for** $h = 1$ to H **do**
2: **for** $k = 1$ to K **do**
3: $AccD[h, k] = LocD[h, k] + \min(AccD[h-1, k], AccD[h-1, k-1], AccD[h, k-1])$



Dynamic Programming

Dynamic Time Warping (DTW) algorithm

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



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 $= 4$

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] $= 0$

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? (5min)

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: $AccD[\mathrm{H,K}] = 0$
Distance ALLDRIG–ALLTID: $AccD[\mathrm{H,K}] = 3$

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

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

Outline

Speech Signal Representations

Template Matching

Probabilistic Approach

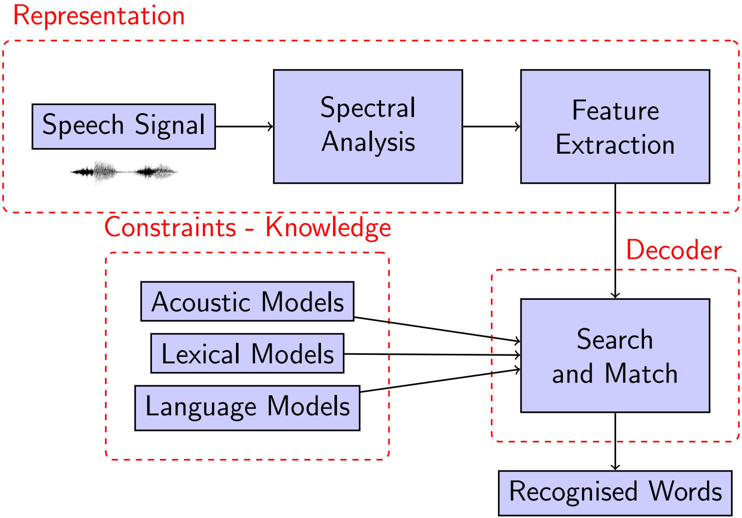
- Knowledge Modelling

Performance Measures

Robustness and Adaptation

Speaker Recognition

Components of ASR System



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:

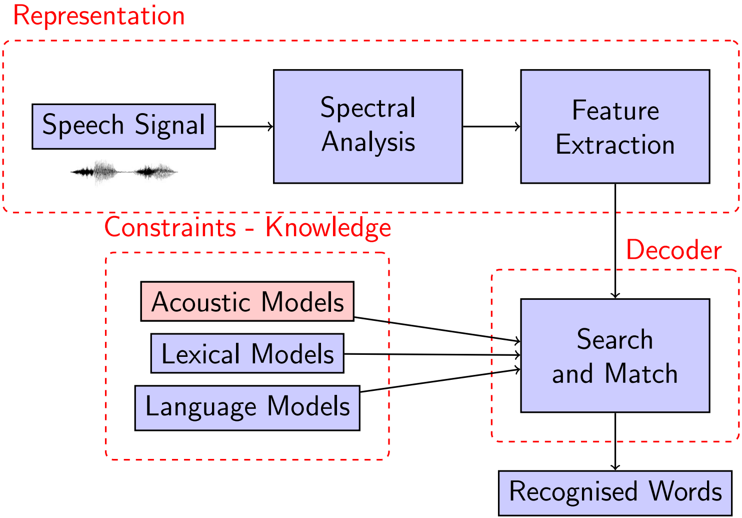
$$\widehat{\text{words}} = \arg\max P(\text{words}|\text{sounds})$$

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)

Components of ASR System

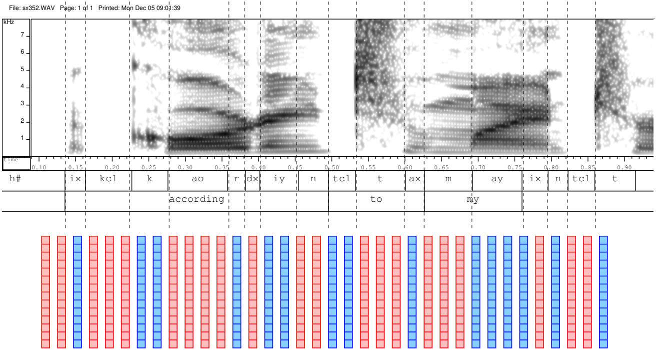


Probabilistic Modelling

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

Probabilistic Modelling

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

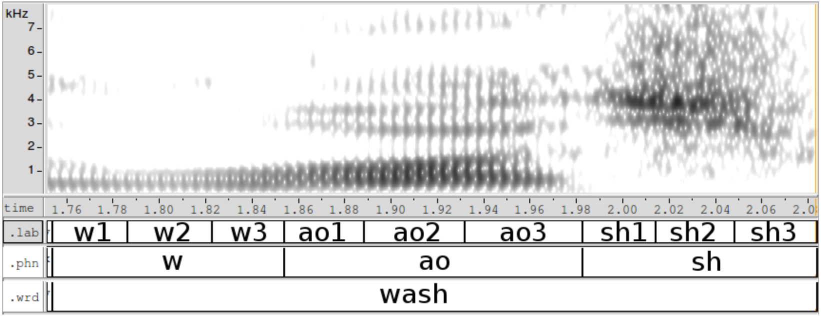


Every feature vector (observation at time t) is a continuous stochastic variable (e.g. MFCC)

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



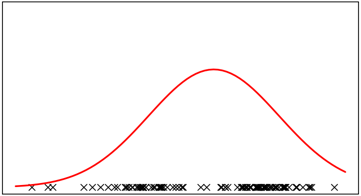
Local probabilities (frame-wise)

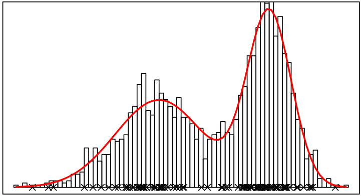
If segment sufficiently short

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

can be modelled with standard probability distributions

Usually Gaussian or Gaussian Mixture

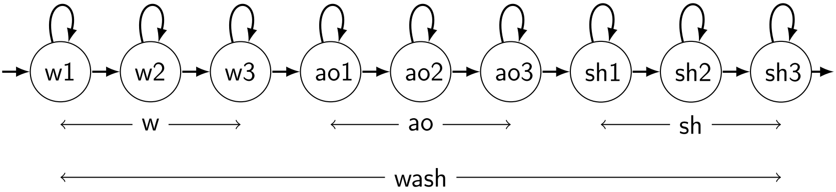




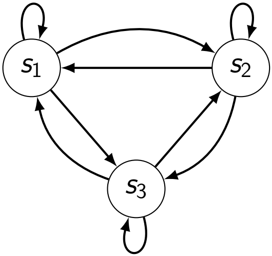
Global Probabilities (utterance)

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

Answer: Hidden Markov Model (HMM)



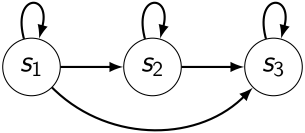
Hidden Markov Models (HMMs)



Elements:

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

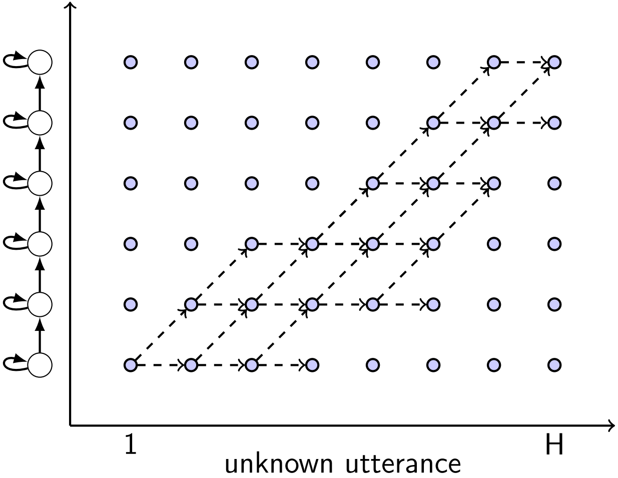
Hidden Markov Models (HMMs)



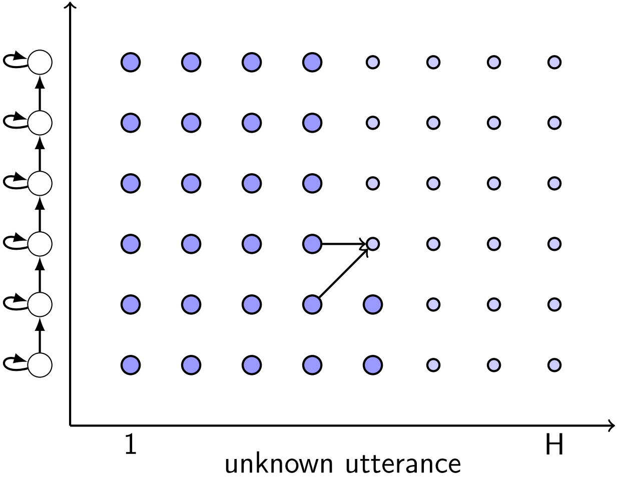
Elements:

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:	$B(o, s_a) = P(o \mid s_a)$

Hidden Markov Models (HMMs)



Hidden Markov Models (HMMs)



HMM-questions

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

HMM-questions

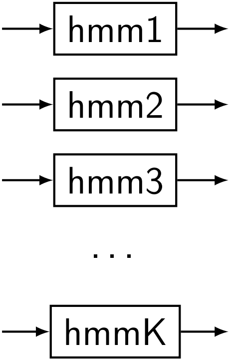
1. what is the probability that the model has generated the sequence of observations? (isolated word recognition) forward algorithm
2. what is the most likely state sequence given the observation sequence? (continuous speech recognition) Viterbi algorithm [5]
3. how can the model parameters be estimated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

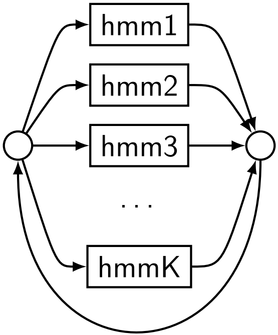
[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". In: *Ann. Math. Statist.* 41.1 (1970), pp. 164–171

Isolated Words Recognition



Compare Likelihoods (forward-backward)

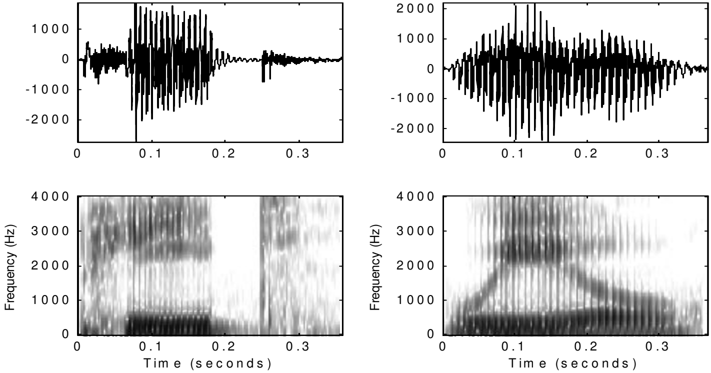
Continuous Speech Recognition



Viterbi algorithm

Modelling Coarticulation

Example peat /piːt/ vs wheel /wiːl/



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

Amount of parameters

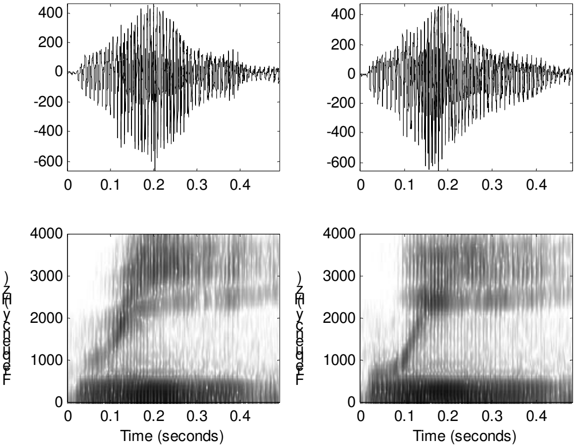
Example:

- 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!

Similar Coarticulation

/riː/ vs /wiː/



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

Tying to reduce complexity

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

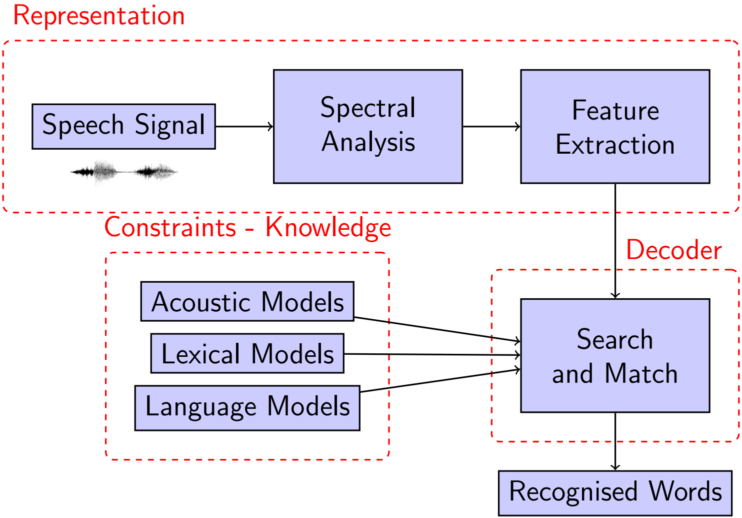
- same right context, similar left context
- 3rd state is expected to be very similar
- 2nd state may also be similar

States (and their parameters) can be shared between models

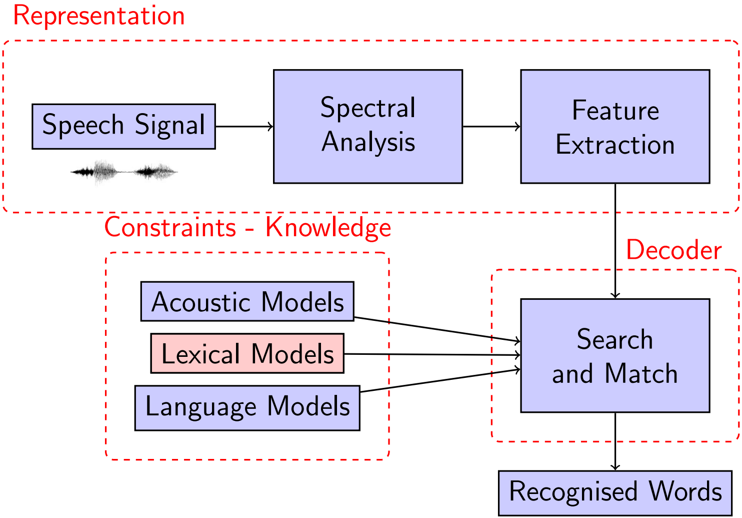
+ reduce complexity
+ more data to estimate each parameter
− fine detail may be lost

done with CART tree methodology

Components of ASR System



Components of ASR System



Lexical Models

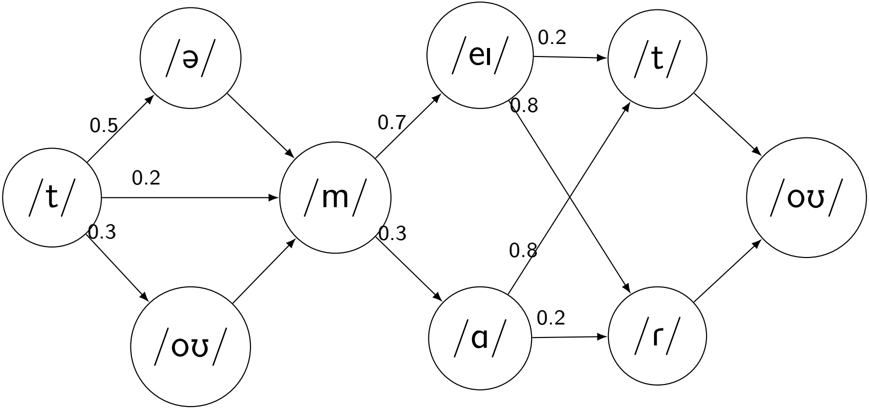
- in general specify sequence of phoneme for each word
- example:

"dictionary"	IPA	X-SAMPA
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)

Pronunciation Network

Example: tomato



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/

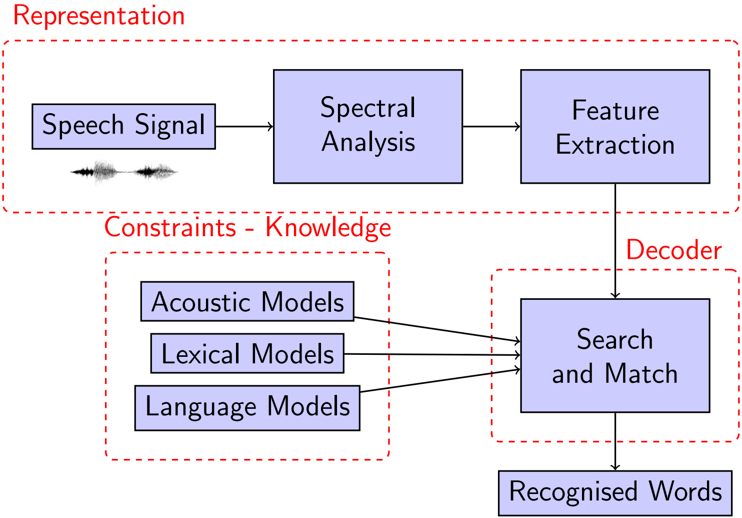
Deletion

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

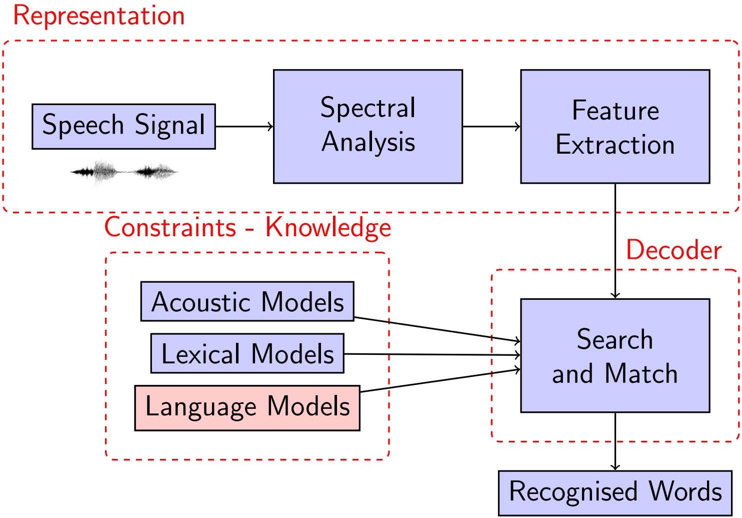
Out of Vocabulary Words

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

Components of ASR System



Components of ASR System



Why do we need language models?

Bayes' rule:

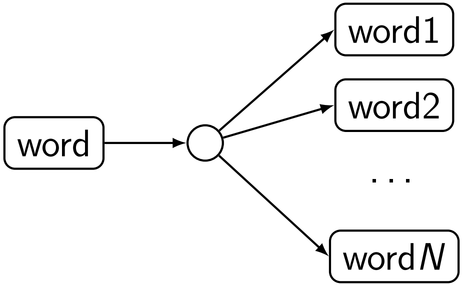
$$P(\text{words}|\text{sounds}) = \frac{P(\text{sounds}|\text{words})P(\text{words})}{P(\text{sounds})}$$

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

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

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



Ambiguity

"ice cream" vs "I scream"

/aɪ s k ɹ iː m/

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

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

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 }

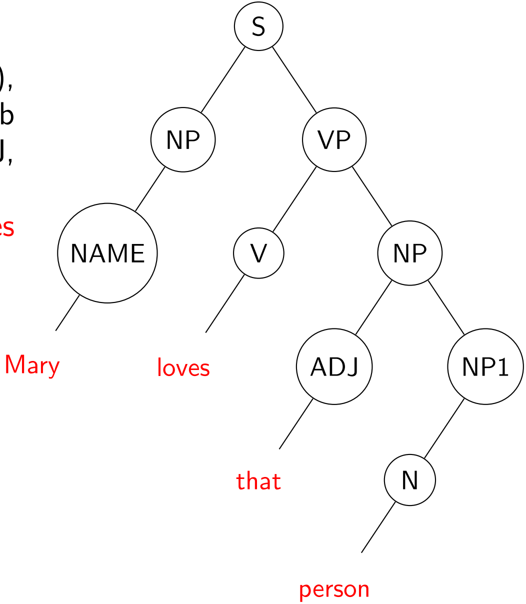
Example

$S =$ sentence

$V =$ {NP (noun phrase), NP1, VP (verb phrase), NAME, ADJ, V (verb), N (noun)}

$T =$ {Mary , person , loves , that , . . . }

$P =$ {S $\to$ NP VP
NP $\to$ NAME
NP $\to$ ADJ NP1
NP1 $\to$ N
VP $\to$ VERB NP
NAME $\to$ Mary
V $\to$ loves
N $\to$ person
ADJ $\to$ that }



Formal Language Models

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

Statistical Grammar Models (N-grams)

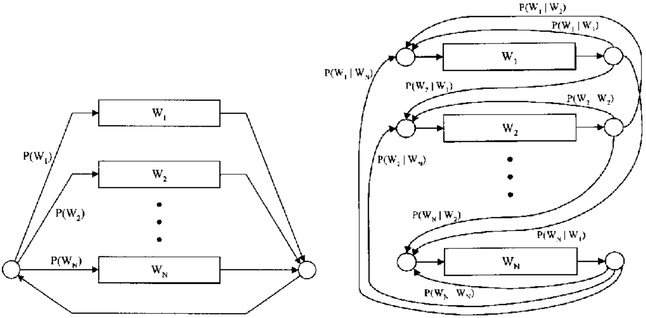
Simply count co-occurrence of words in large text data sets

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

Language Models: complexity

Increasing N in N-grams leads to:

1. more complex decoders



2. difficulties in training the LM parameters

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)

Outline

Word Accuracy

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

Where

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

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

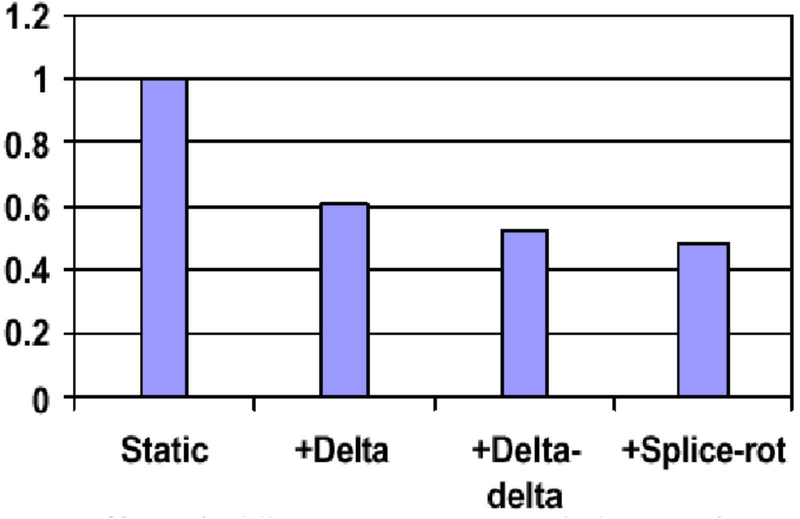
Measure Difficulty

Language Perplexity

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

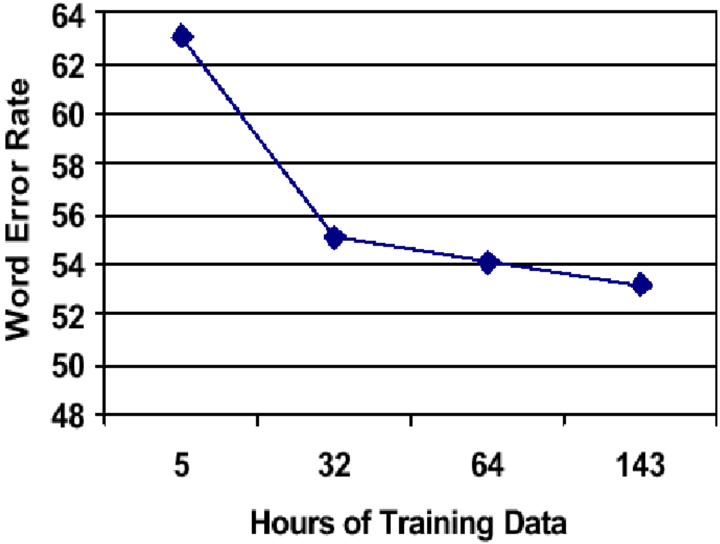
- $P(W)$ is the probability of the word sequence (language model)
- H is called entropy
- B can be seen as measure of average number of words that can follow any given word
- Example: equiprobable digit sequences $B = 10$

Effect of adding features

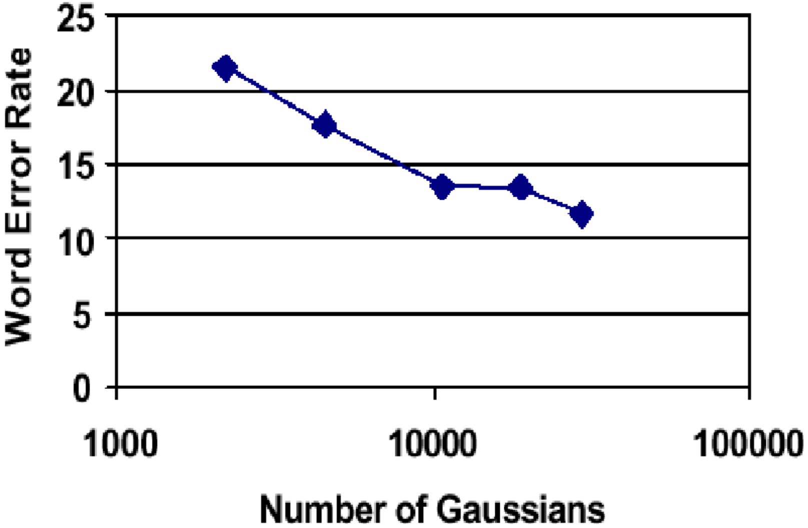


Effect of adding training data

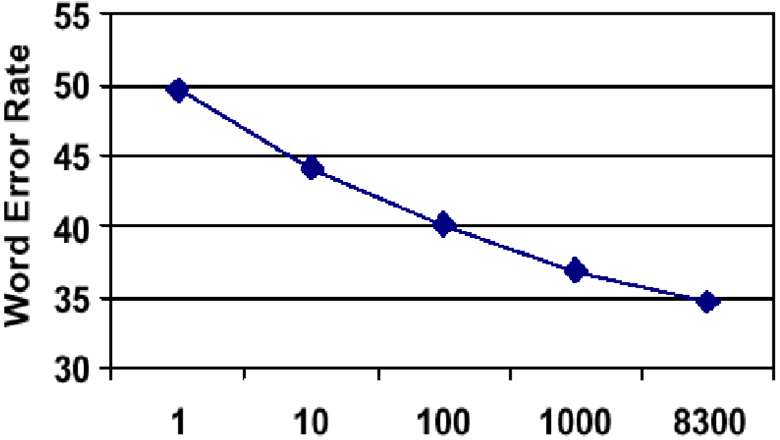
Swichboard data



Effect of adding Gaussians



Effect of adding data for language models



Million sentences

Some dictation systems

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

New applications

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

Outline

Main variables in ASR

Speaking mode isolated words vs continuous speech

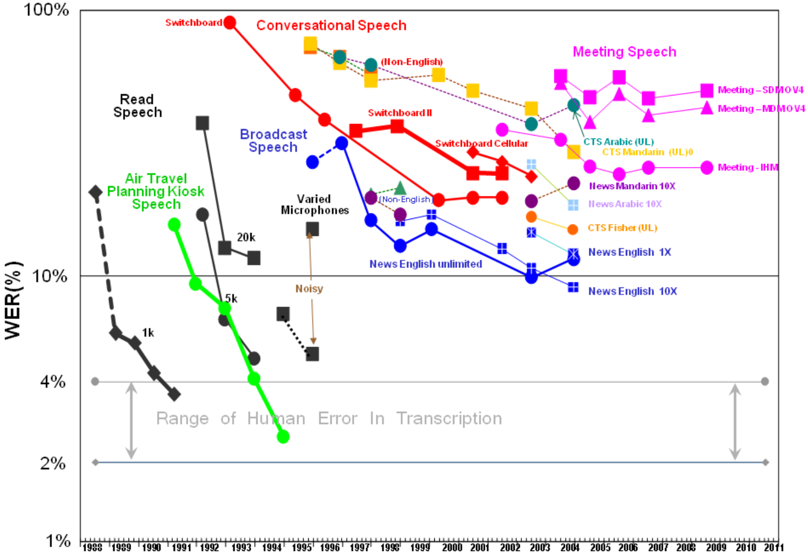
Speaking style read speech vs spontaneous speech

Speakers speaker dependent vs speaker independent

Vocabulary small ($<$20 words) vs large ($>$50 000 words)

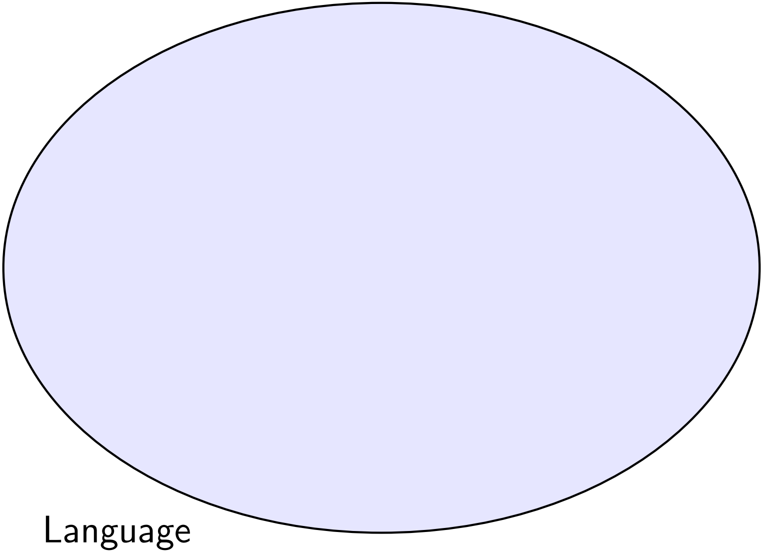
Robustness against background noise

NIST STT Benchmark Test History – May. '09

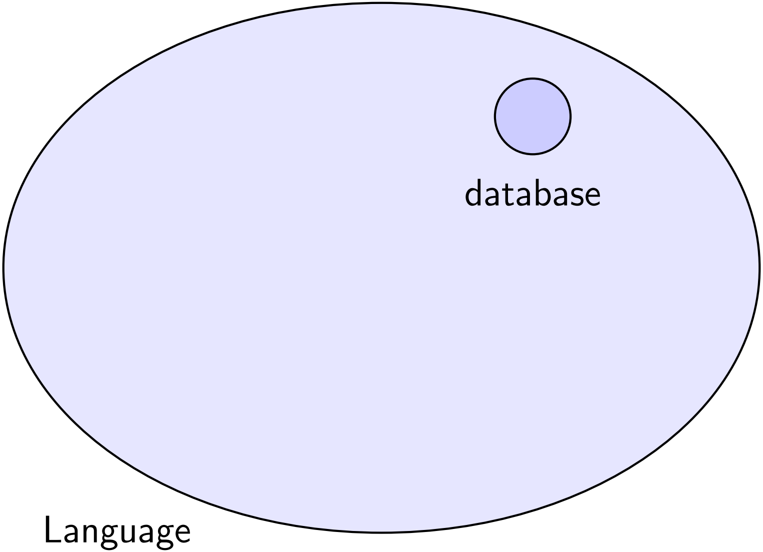


http://www.itl.nist.gov/iad/mig/publications/ASRhistory/

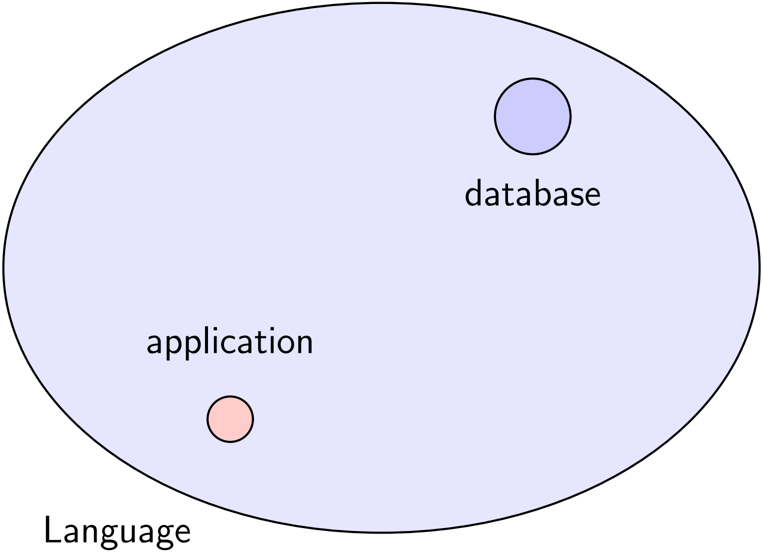
Why is it so hard?



Why is it so hard?



Why is it so hard?



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

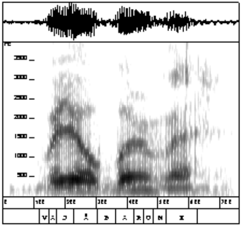
Sheep and Goats [3]

[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

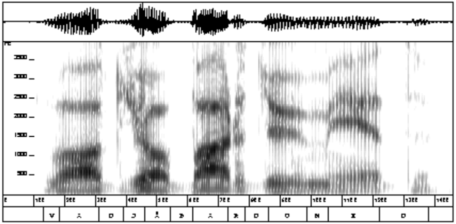
Sheep and Goats [3]

[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

Exmpl: spontaneous vs hyper-articulated



Va jobbaru me



Vad jobbar du med

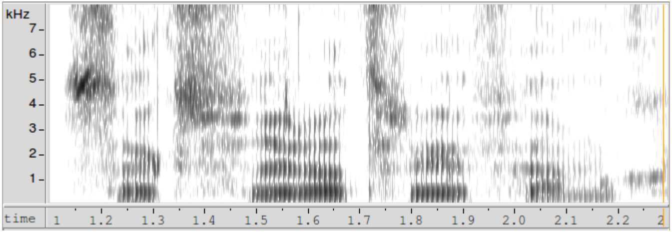
“What is your occupation”
(“What work you with”)

Examples of reduced pronunciation

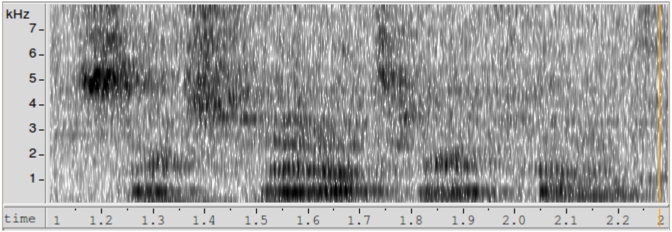
Spoken	Written	In English
Tesempel	Till exempel	for example
åhamba	och han bara	and he just
bafatt	bara för att	just because
javende	jag vet inte	I don't know

Microphone distance

Headset

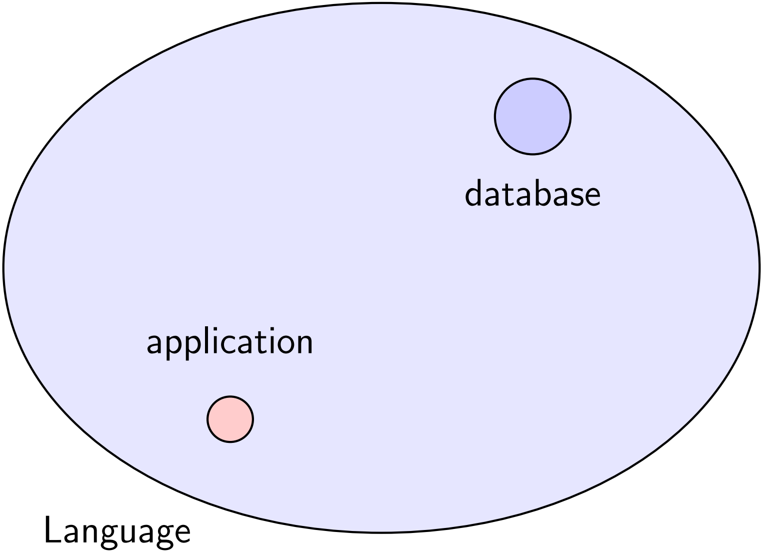


2 m distance



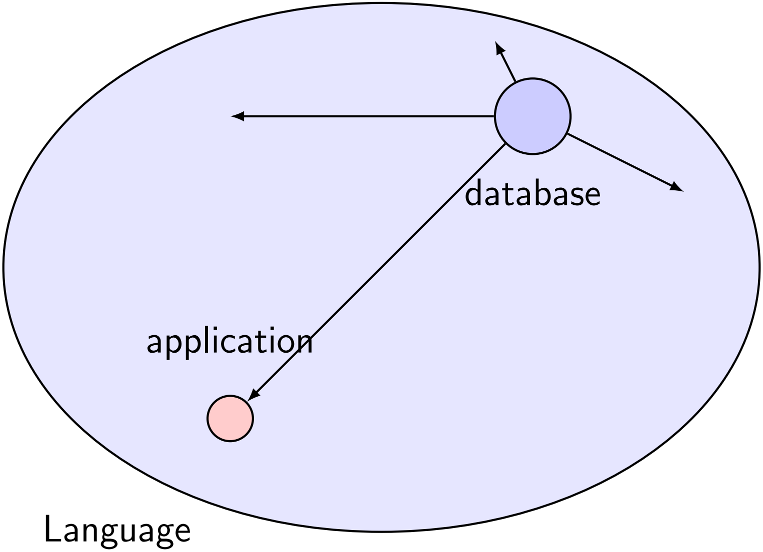
How do we cope with variability?

Ideally: models that generalise



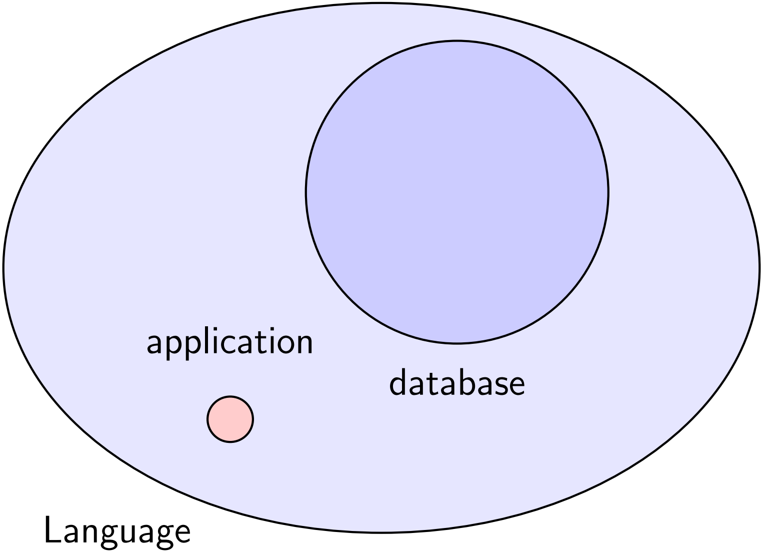
How do we cope with variability?

Ideally: models that generalise



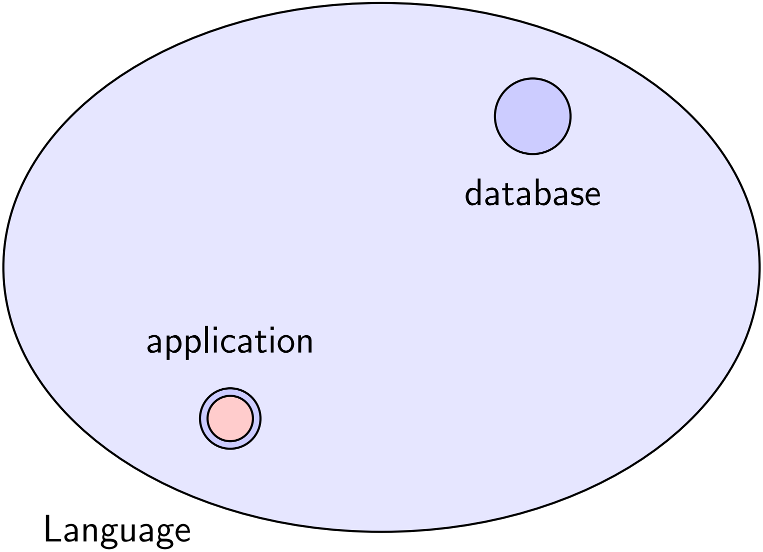
How do we cope with variability?

Large companies use insane quantities of data



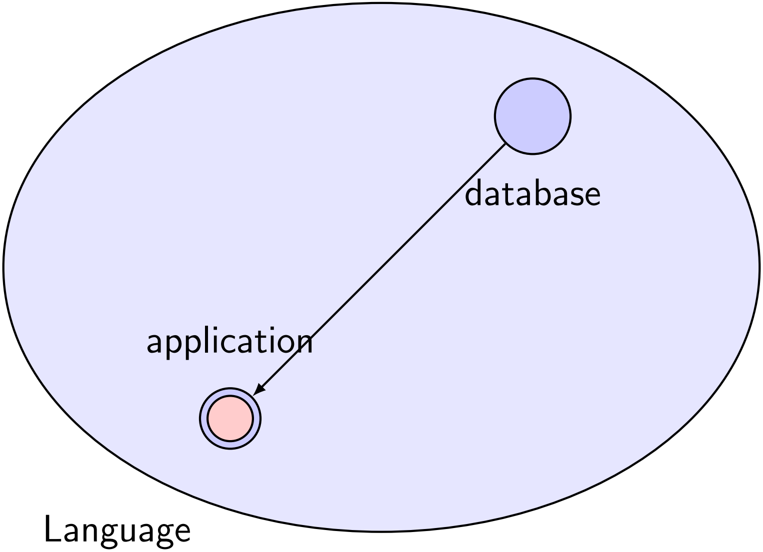
How do we cope with variability?

Adaptation



How do we cope with variability?

Adaptation



Adaptation: Example

Enrolment in Dictation Systems

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

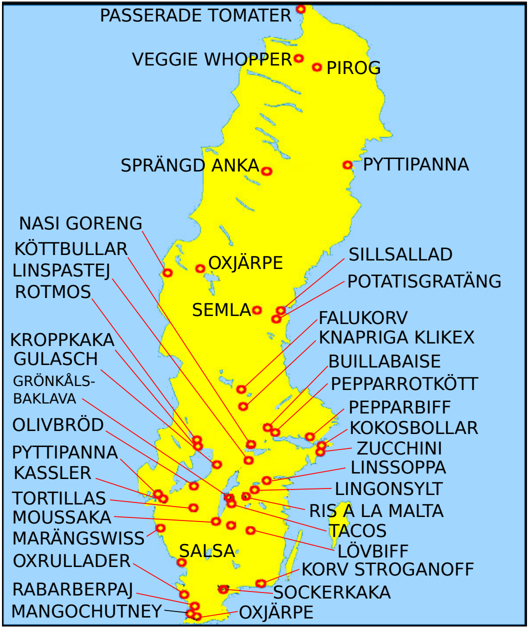
Beta version of smartphone applications

- the company has all the rights on data generated

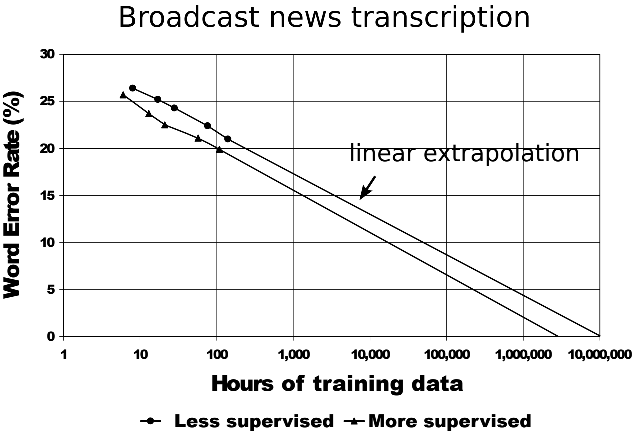
Limitations

- lack of context
- require huge amounts of training data

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



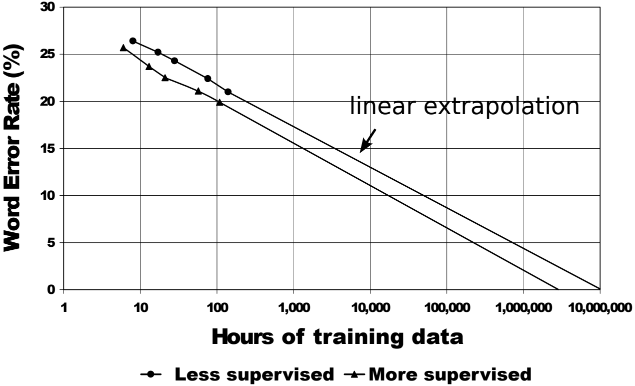
Lack of Generalisation[4]



[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

Lack of Generalisation[4]

Broadcast news transcription



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

---

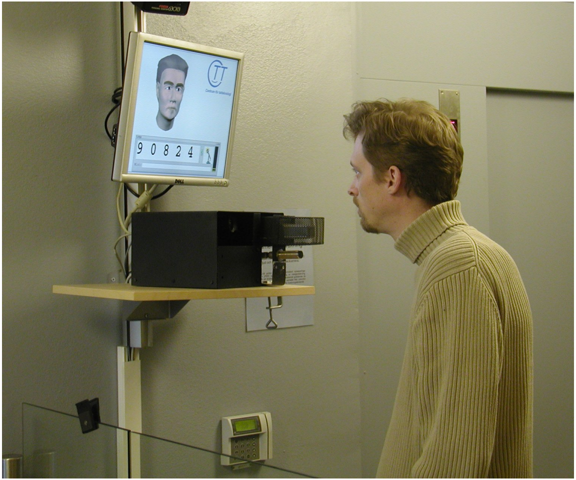
[4] R. Moore. "A Comparison of the Data Requirements of Automatic Speech Recognition Systems and Human

New directions

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

Outline

Speaker Recognition



Created by Håkan Melin

Person Identification

Methods rely on:

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

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

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

Representations

Speech Recognition:

- represent speech content
- disregard speaker identity

Speaker Recognition:

- represent speaker identity
- disregard speech content

Representations

Speech Recognition:

- represent speech content
- disregard speaker identity

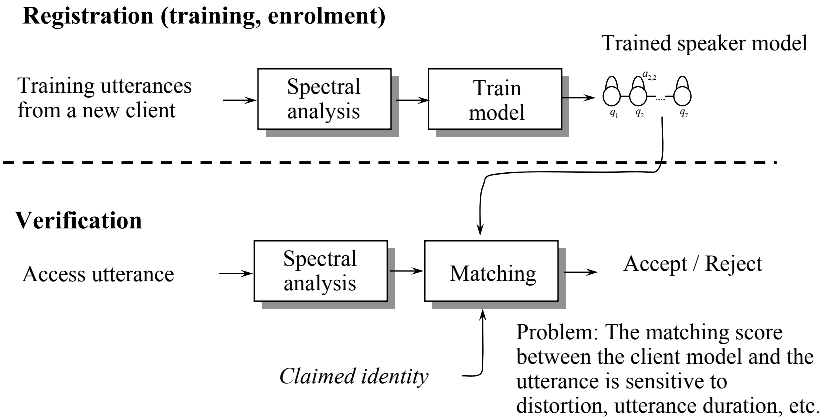
Speaker Recognition:

- represent speaker identity
- disregard speech content

Surprisingly:

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

Speaker Verification



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

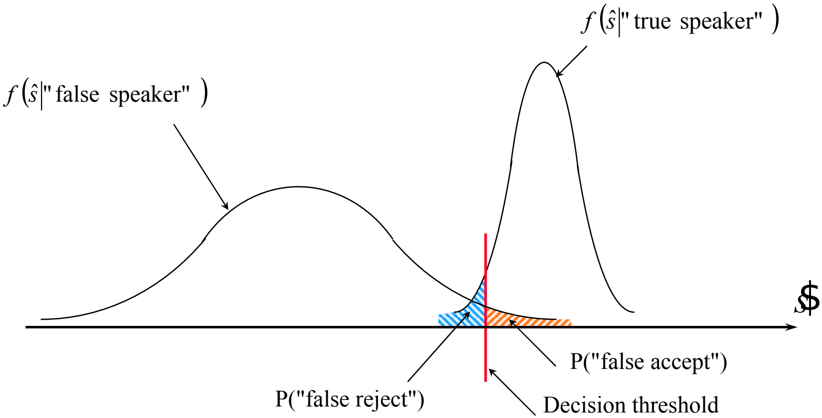
SVM (Support Vector Machines)

Combined models

Evaluation

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

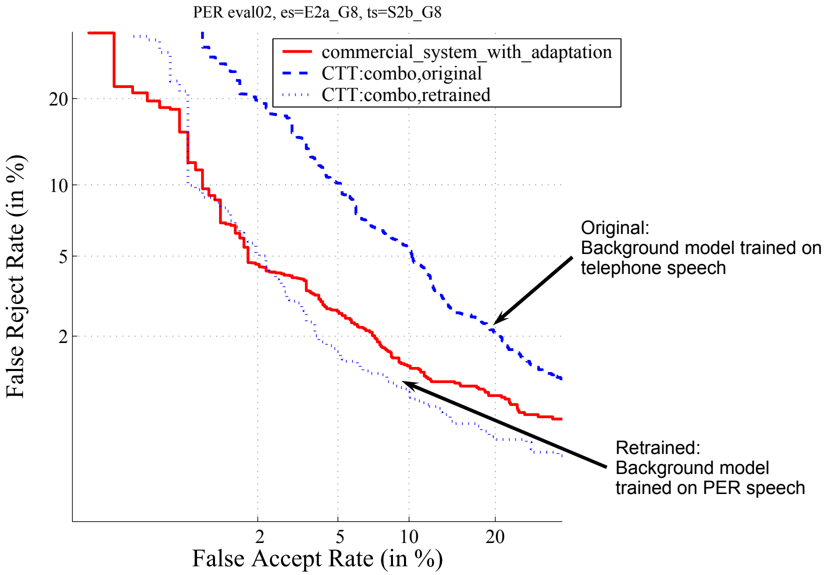
Score Distribution and Error Balance



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

PER vs Commercial System



More information and mathematical formulations in DT2118