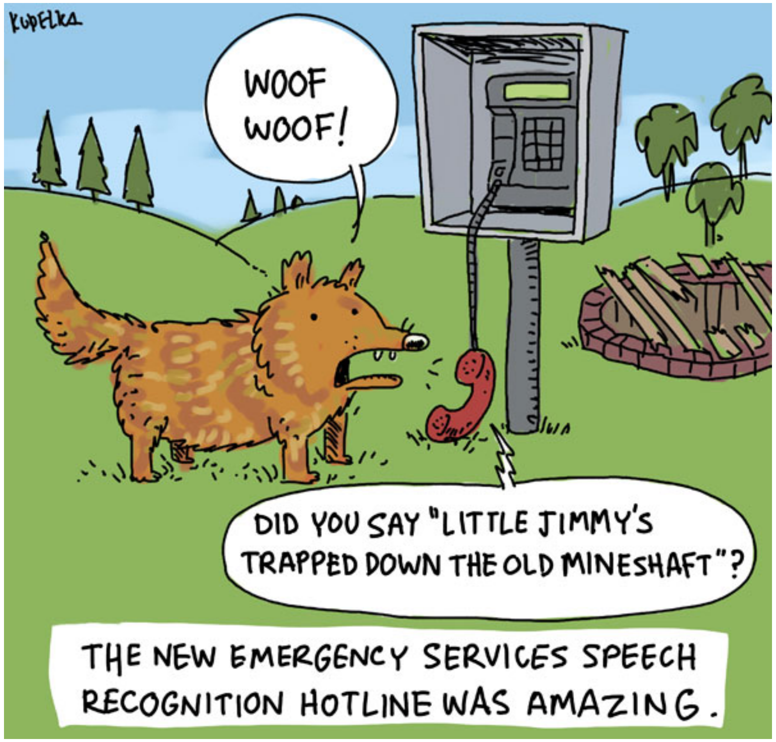
# Automatic speech recognition


Dr. Odette Scharenborg
Multimedia Computing Group
Intelligent Systems / Computer Science

OScharenborg


TUDelft

# Learning objectives

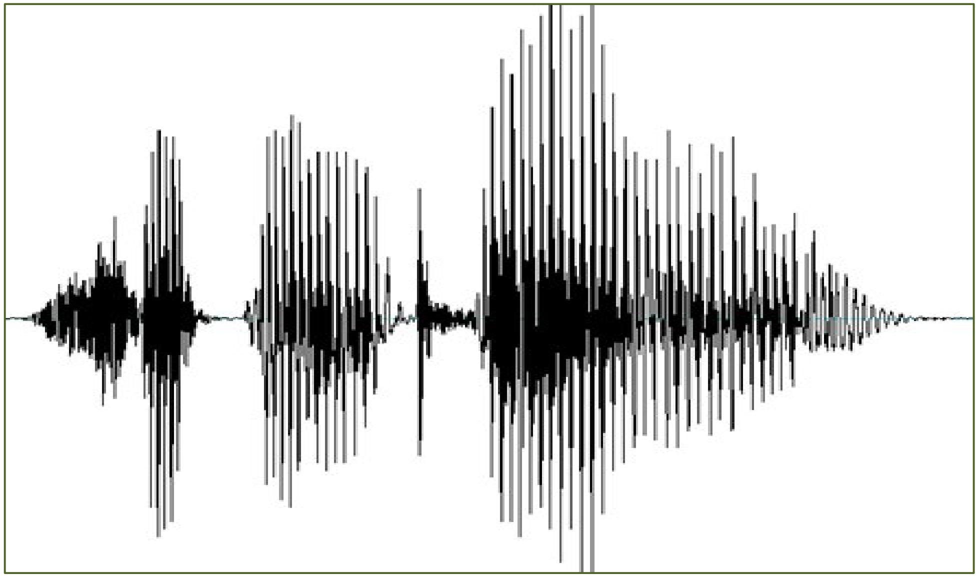
After today you will

- Be able to explain what speech is
- Know the goal, basic architecture and workings of an automatic speech recognition (ASR) system
- Know how ASR systems are evaluated and how well they perform
- The limitations of an ASR

# What is speech?

# Speech

- Speech = sound = differences in air pressure
- Perceived as different phone(me)s, phone(me) sequences, words



speech signal

# Some terminology

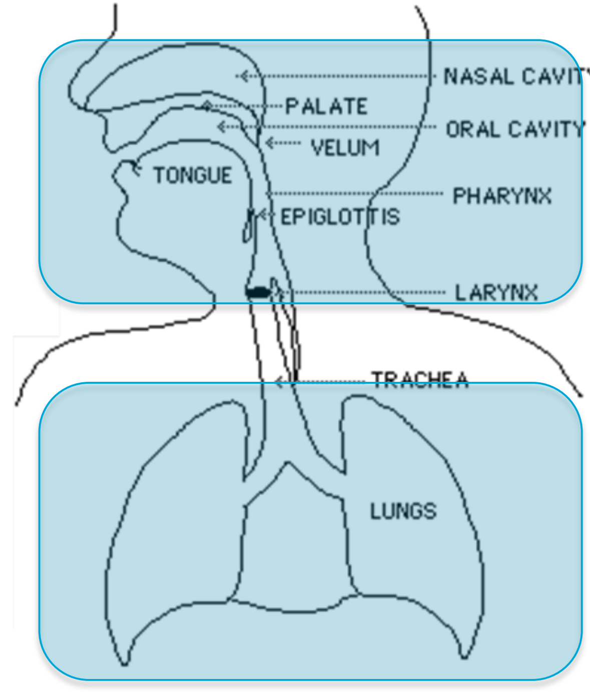
- Words: sequences of phonemes
- Phoneme: the smallest contrastive linguistic unit that distinguishes meaning, e.g., *tip* vs. *dip*
- Allophone: a variation of a phoneme, e.g., $p^h$*ot* vs. *spot*
- Phone: a distinct speech sound

TU Delft

# The speech production system

Vocal tract

- Area between vocal cords and lips
- Pharynx + nasal cavity + oral cavity



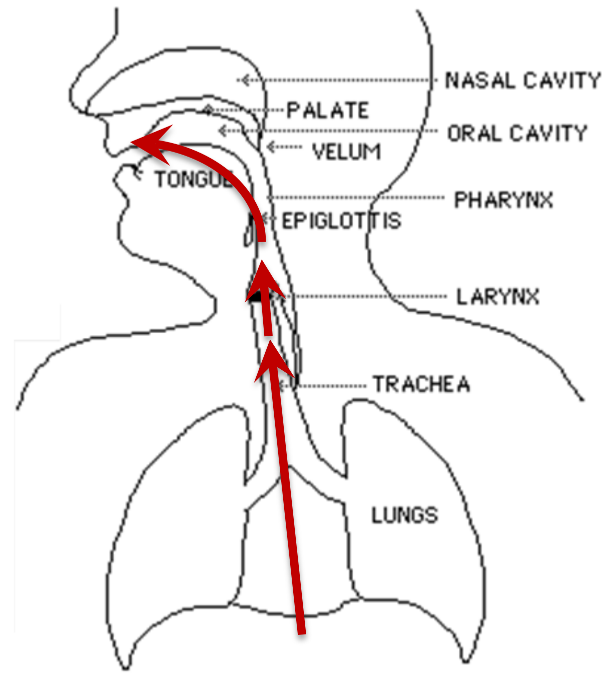

and lungs

# 3 steps to produce sounds

step 3: *articulation* =
distortion of air
= speech

step 2: *phonation*

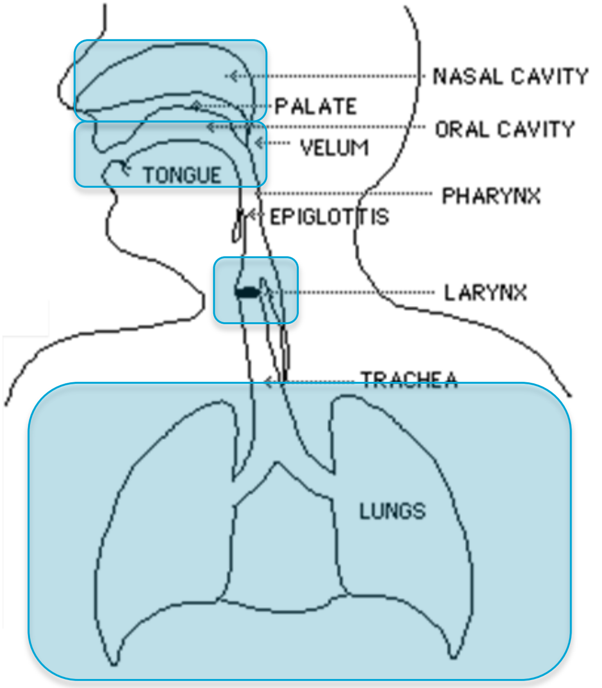
step 1: *initiation*




TU Delft

# Fun fact

None of the speech production components are specifically made for speaking!



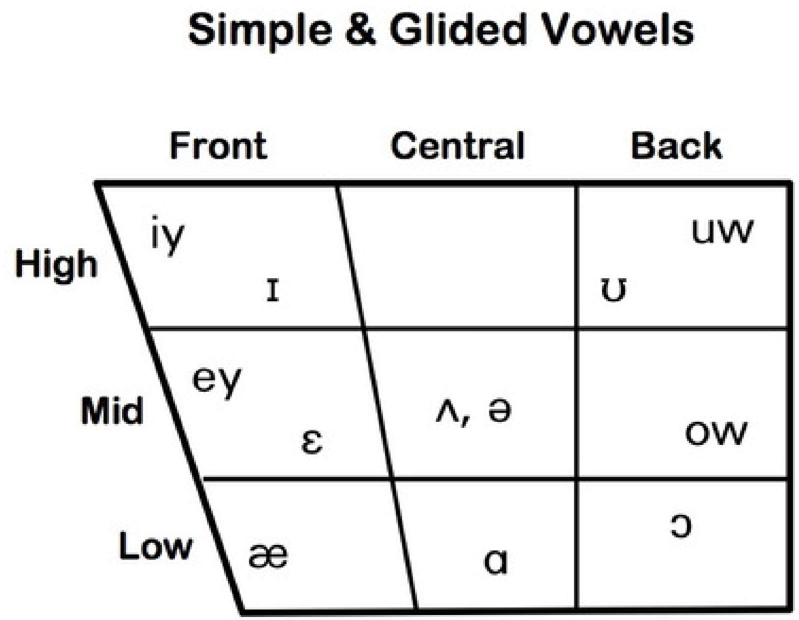

TU Delft

# Speech sounds

- Vowels: unblocked air stream
- Consonants: constricted or blocked air stream

# Different sounds: Vowels

- Tongue height:
  - Low: e.g., /a/
  - Mid: e.g., /e/
  - High: e.g., /i/
- Tongue advancement:
  - Front : e.g., /i/
  - Central : e.g., /ə/
  - Back : e.g., /u/
- Lip rounding:
  - Unrounded: e.g., /ɪ, ɛ, e, ə/
  - Rounded: e.g., /u, o, ɔ/



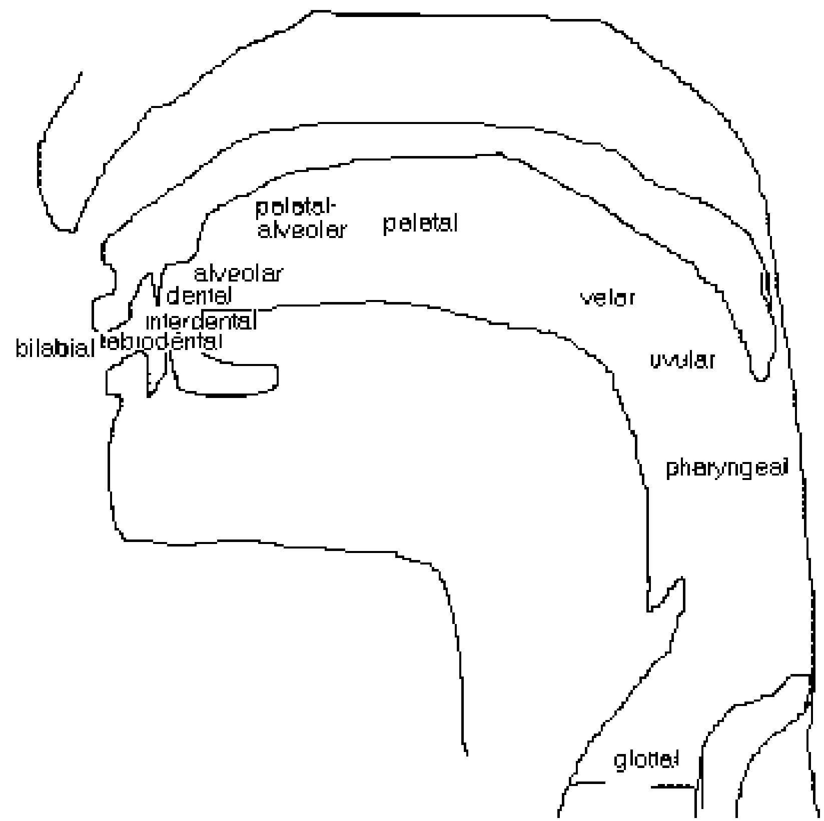
*English vowel quadrant*

# Different sounds: Consonants

- **Place of articulation**
  Where is the constriction?

- **Manner of articulation**
  – Stops: /p, t, k, b, d, g/
  – Fricatives: /f, s, S, v, z, Z/
  – Affricates: /tS, dZ/
  – Approximants/Liquids: /l, r, w, j
  – Nasals: /m, n, ng/

- **Voicing**



# Speech sound production

- https://www.youtube.com/watch?v=DcNMCB-Gsn8

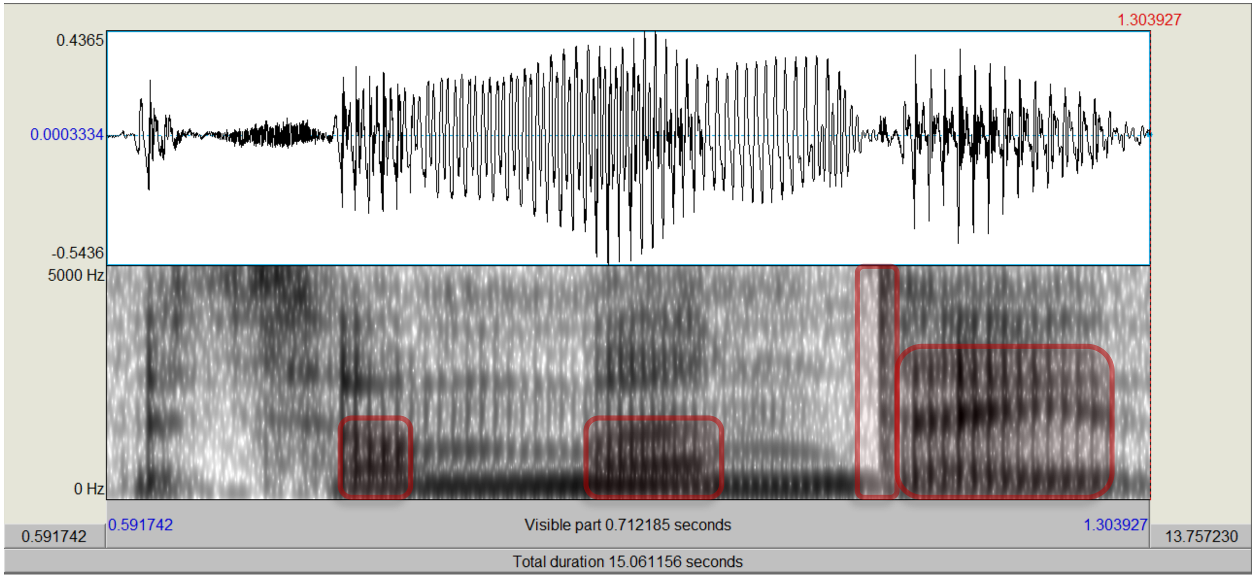
*Recorded in 1962, Ken Stevens*
*Source: YouTube*

The physical speech signal consists of

… acoustic energy

… varying over time in amplitude and spectral shape

# Each sound has its own spectral shape




bu t o nM o n d ay

# Demos of speech sound manipulations

- http://jontalle.web.engr.illinois.edu/Public/InterspeechDemosAug25.13/da_to_ga_f103.m4v

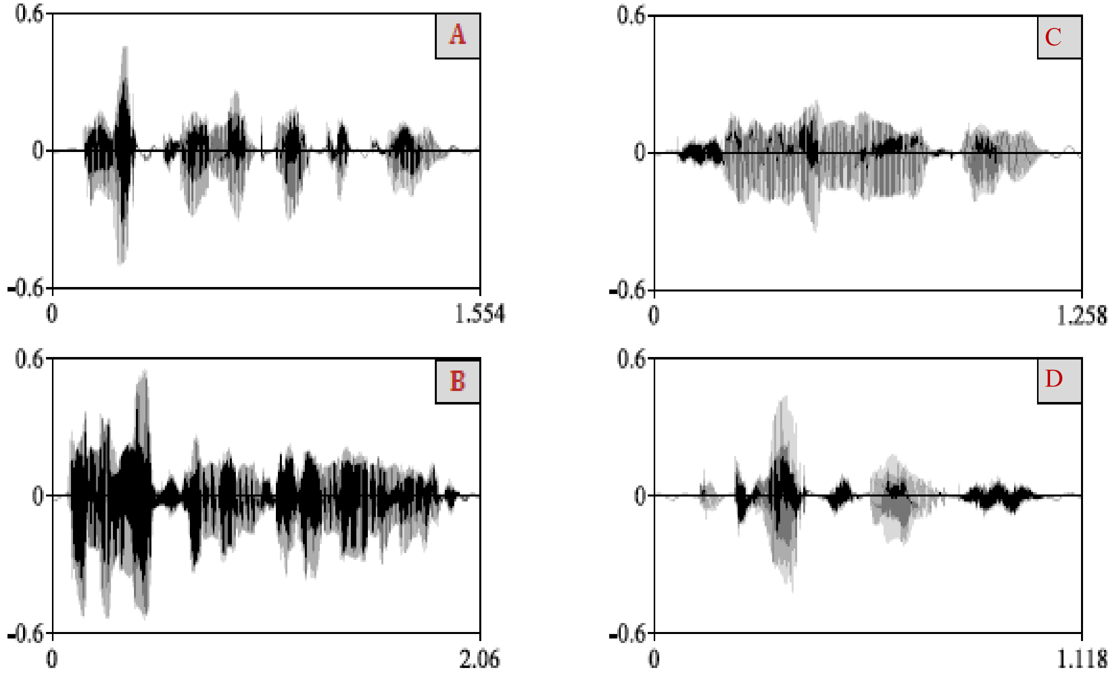
- http://jontalle.web.engr.illinois.edu/Public/InterspeechDemosAug25.13/ka_to_ta_f103.m4v

- http://jontalle.web.engr.illinois.edu/Public/InterspeechDemosAug25.13/Sa2sa2cha2za2Da.m4v

# 3 important aspects of speech

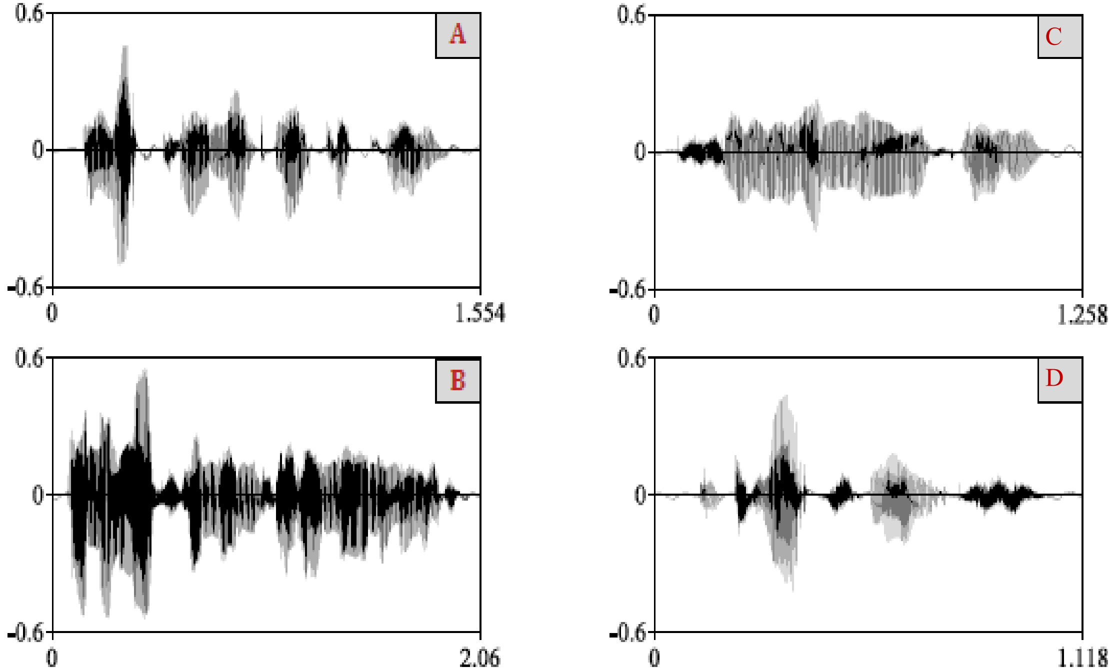
# Quiz 1: Count the words

Each picture shows a waveform of a short stretch of speech:



# Quiz 1: Count the words

Each picture shows a waveform of a short stretch of speech:




A: Electromagnetically (1)

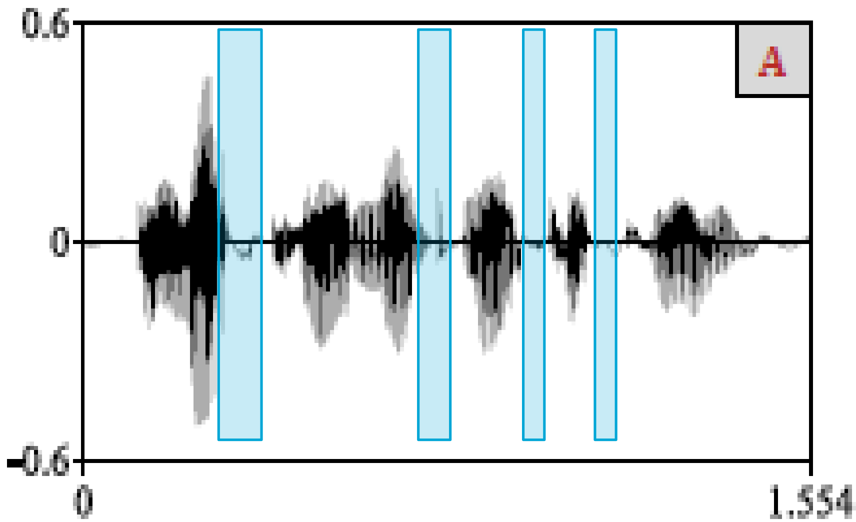
B: Emma loves her mum's yellow marmelade (6)

C: See you in the evening (5)

D: Attachment (1)

# Electromagnetically

Why is it so hard to determine the number of words?

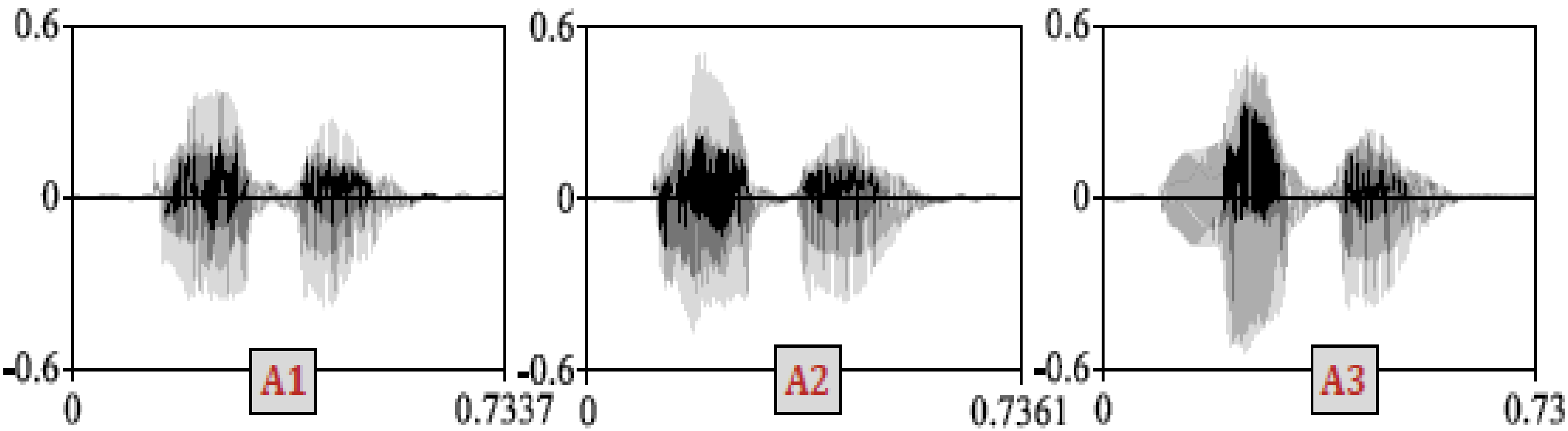



/ı l ɛktromæ g nɛt ɪ k əll/

silence ≠ word boundary

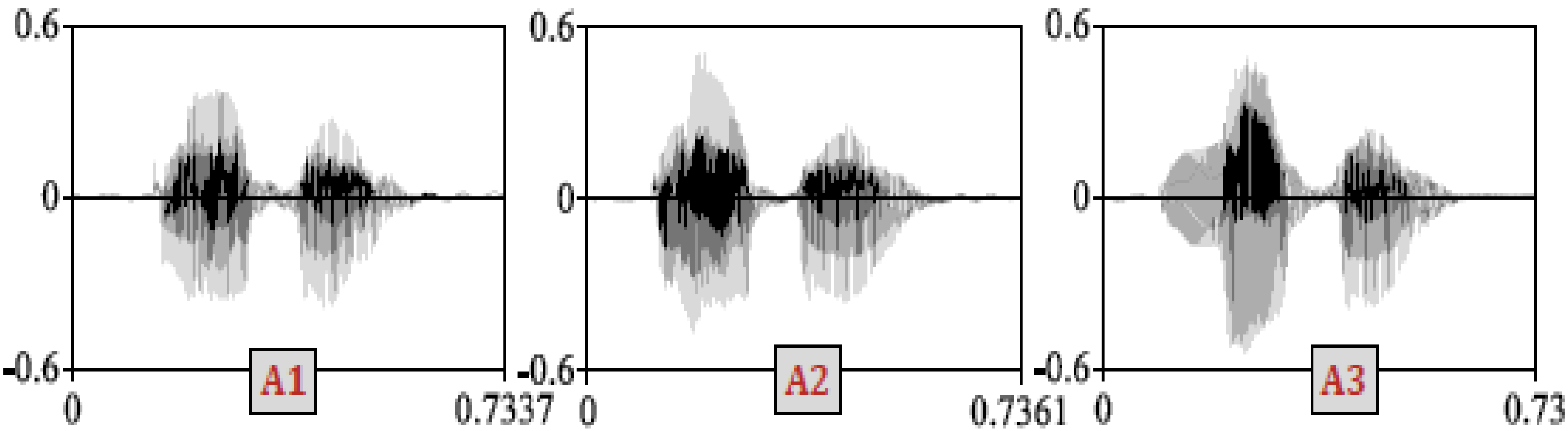
# Quiz 2: Spot the odd one out

- Below are three waveforms each containing a single word:



# Quiz 2: Spot the odd one out

- Below are three waveforms each containing a single word:




*Every time you produce a word it sounds differently*

A3 (brother, brother, mother)

# Enormous variability

**Speaker-dependent:**

- Speaker differences, e.g., gender, vocal tract length, age
- Speaker idiosyncracies, e.g., lisp, creaky voice
- Accent: dialects, non-nativeness

**Speaker-independent:**

- Background noise

# Enormous variability

**Speaker-independent:**

- Coarticulation: production of a speech sound becomes more like that of a preceding/following speech sound, e.g.
  - Place of articulation: garden bench → gardem bench (*anticipatory* or *regressive* coarticulation)
  - Voicing: cats vs. dogz (*carryover* coarticulation)

- Speaking style
  - Formal
  - Read
  - Informal, conversational → reductions

TU Delft

# Reductions

**natuurlijk** (of course)

/natyrl@k/

/naty l@k/

/ ty l@k/

/ ty k/

**eigenlijk** (actually)

/Eix@nl@k/

/Eix@ l@k/

/Eix l@k/

/Ei k/

TU Delft

# Summary of 3 important aspects

- Speech signal is continuous
- No clear pauses between words
- Highly variable

> Task for the ASR system:
> Map the highly variable, continuous speech signal onto discrete units such as words

# Automatic speech recognition

# Automatic speech recognition

**Task:** Automatic conversion of the speech signal into a

- Input = ordered, time-continuous sequence
- Output = ordered text sequence

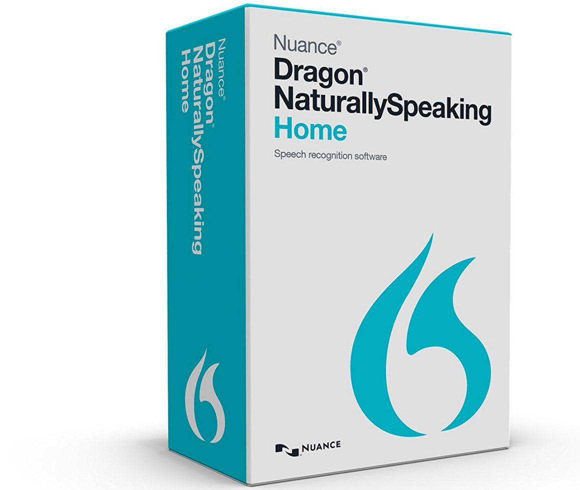
**Goal:** Do this under a variety of listening and speaker conditions, with the least possible number of recognition errors

TU Delft

# Related tasks

- Speech understanding: generating a semantic representation
- Speaker recognition: identifying the person who spoke
- Speech detection: separating speech from non-speech
- Speech enhancement: improve the intelligibility of a signal
- Speech compression: encode speech signal for transmission or storage with a small amount of bits

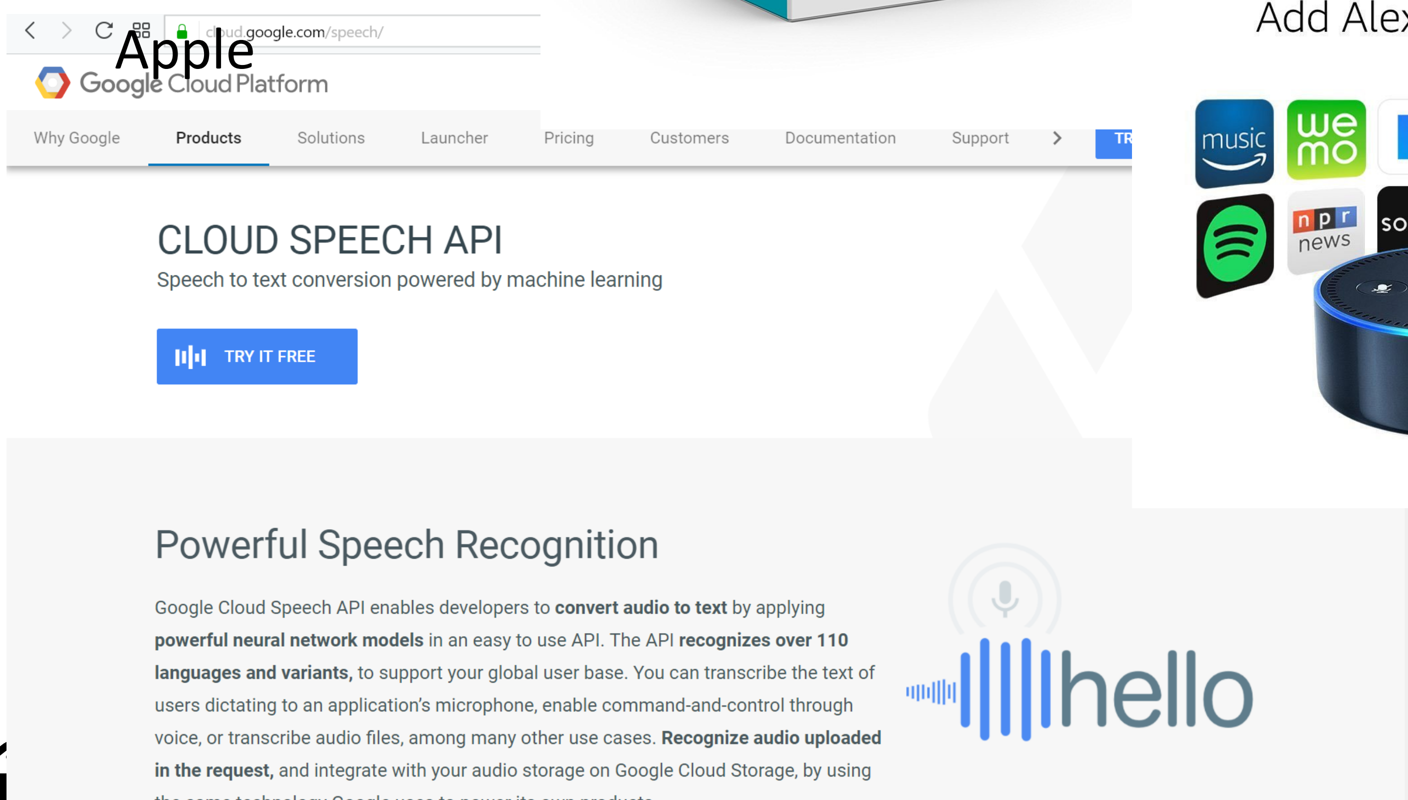
Siri - Apple




Google
ASSISTANT

echo dot

Add Alexa to any room



cloud.google.com/speech/

Google Cloud Platform

Why Google | Products | Solutions | Launcher | Pricing | Customers | Documentation | Support

# CLOUD SPEECH API

Speech to text conversion powered by machine learning

TRY IT FREE

## Powerful Speech Recognition

Google Cloud Speech API enables developers to **convert audio to text** by applying **powerful neural network models** in an easy to use API. The API **recognizes over 110 languages and variants,** to support your global user base. You can transcribe the text of users dictating to an application's microphone, enable command-and-control through voice, or transcribe audio files, among many other use cases. **Recognize audio uploaded in the request,** and integrate with your audio storage on Google Cloud Storage, by using the same technology Google uses to power its own products.

hello

# The noisy channel

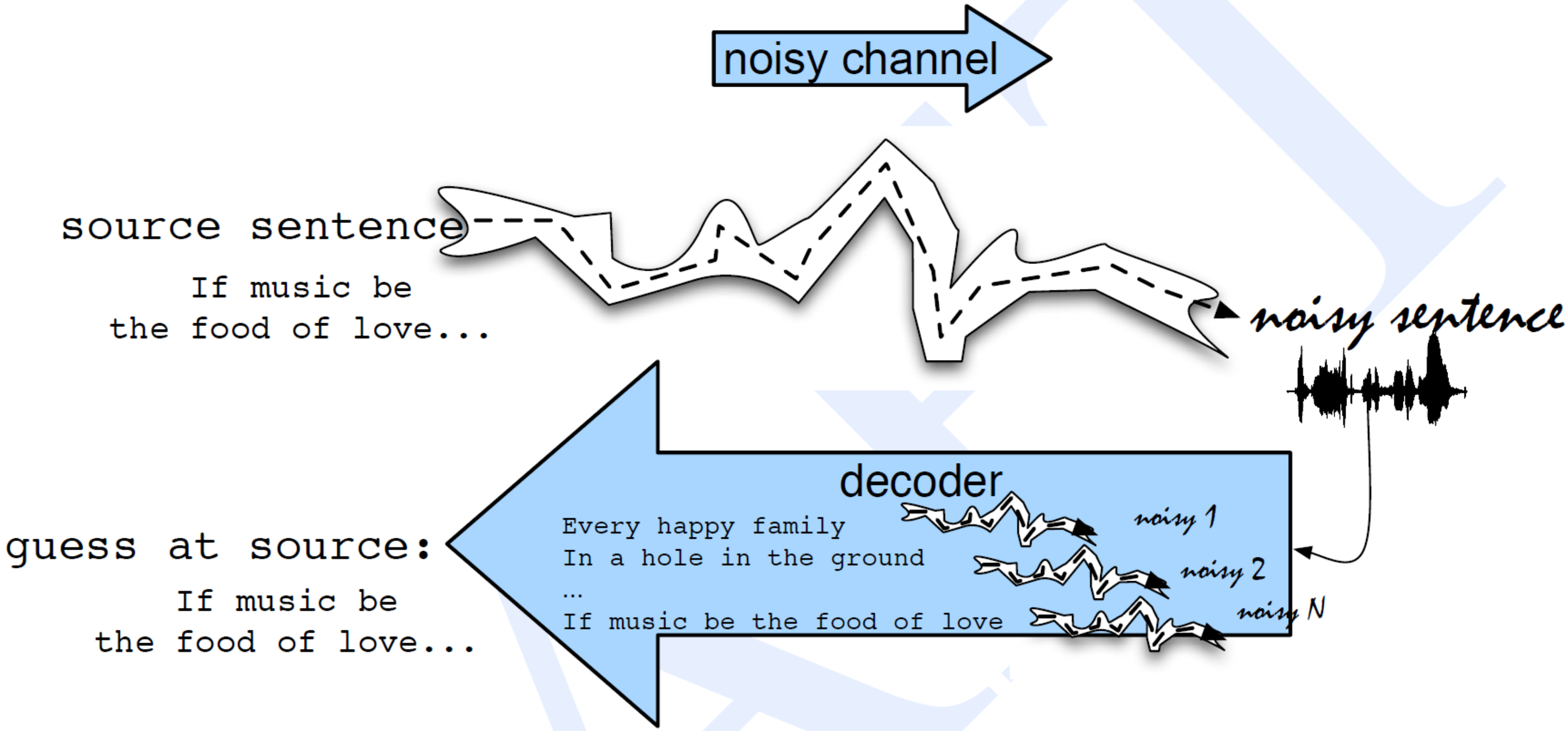



*Figure from Jurafsky and Martin, edition 2*

# Two big problems

- Speech is highly variable, will never exactly match any model we have for the sentence
- We need a metric to determine the "best match"

→ Probability → Bayesian inference

- Set of all sentences is huge
- We need an efficient algorithm that does not search through all possible sentences but only the most likely ones

→ Search or decoding problem

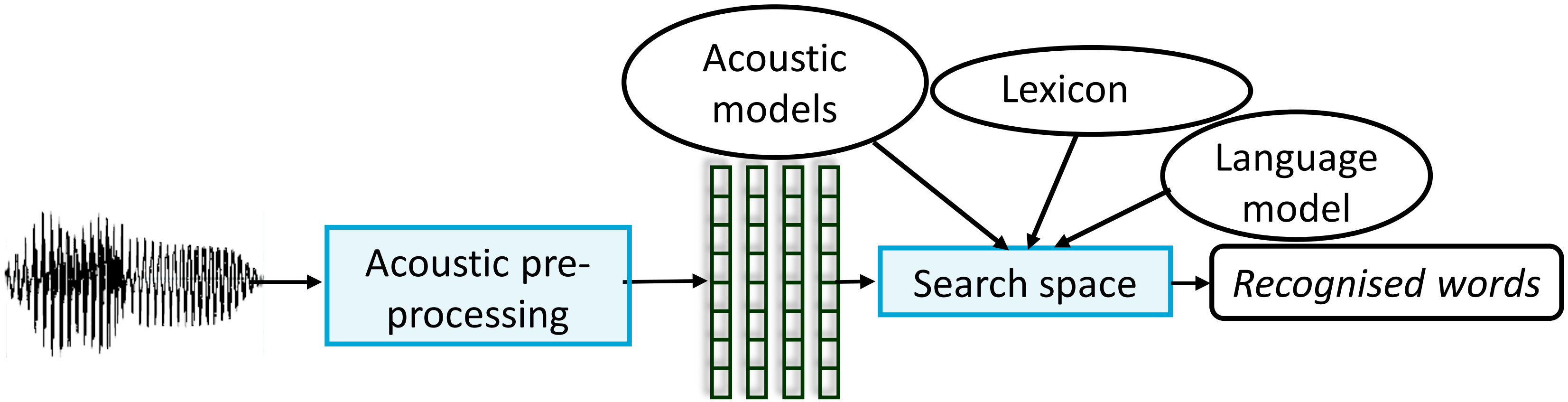
# Goal ASR

Find the **most likely sentence *W* out of all sentences** in the language *L* given some acoustic input *X*

Bayes Theorem: $$\text{argmax } P(W|X) = \frac{P(X|W) \cdot P(W)}{\cancel{P(X)}}$$

TUDelft

# ASR system




Speech recognition is the problem of deciding on

- How to *represent* the signal
- How to *model* the constraints (P(X|W) and P(W))
- How to *search* for the most optimal answer (P(W|X))

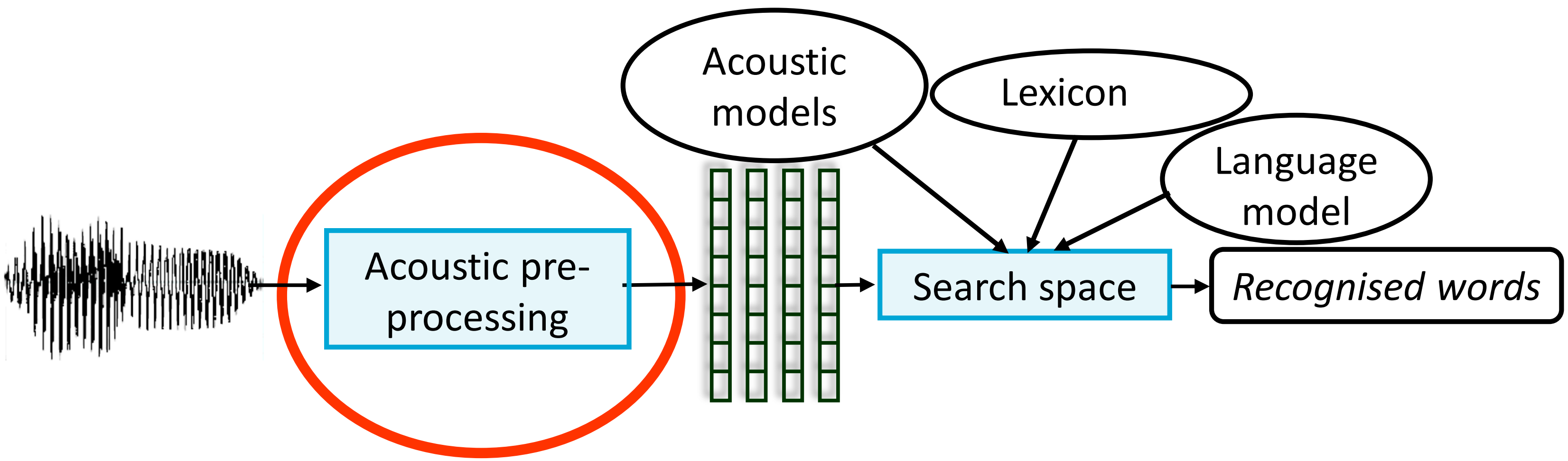
Slide partially based on slide by James Glass, MIT

# How to represent the speech signal

# Acoustic pre-processing

= Computation of acoustic feature vectors of the speech signal

➔ Mel-frequency cepstral coefficients



# How to model the constraints

1. Acoustic model: to model the constraints inherent to the speech signal

2. Lexicon: to model the constraints on the order of sounds in a language
   → Determined by the words of a language

3. Language model: to model the constraint inherent to word order in the language

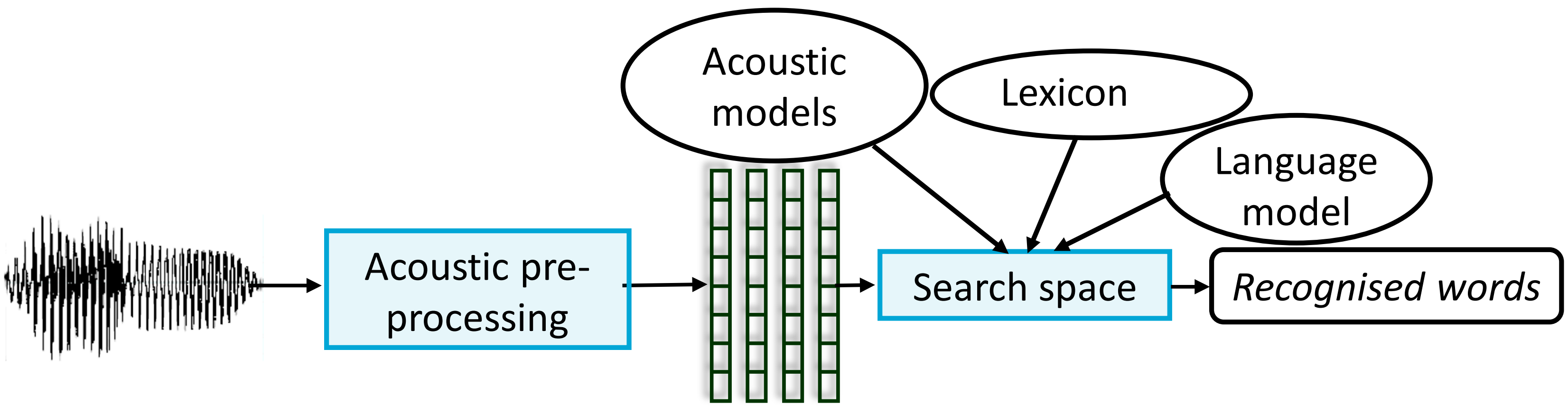
TU Delft

# How to obtain P(X|W)?

- Derive an estimate of the probability that a particular recognition unit *W* generated a particular stretch of speech *X*

→ $P(X|W)$

- *P* = probability
- *W* = word
- *X* = sequence of acoustic vectors (typically MFCCs)




Bayes Theorem: argmax P(W|X) = ( P(X|W) • P(W) )

# Acoustic models

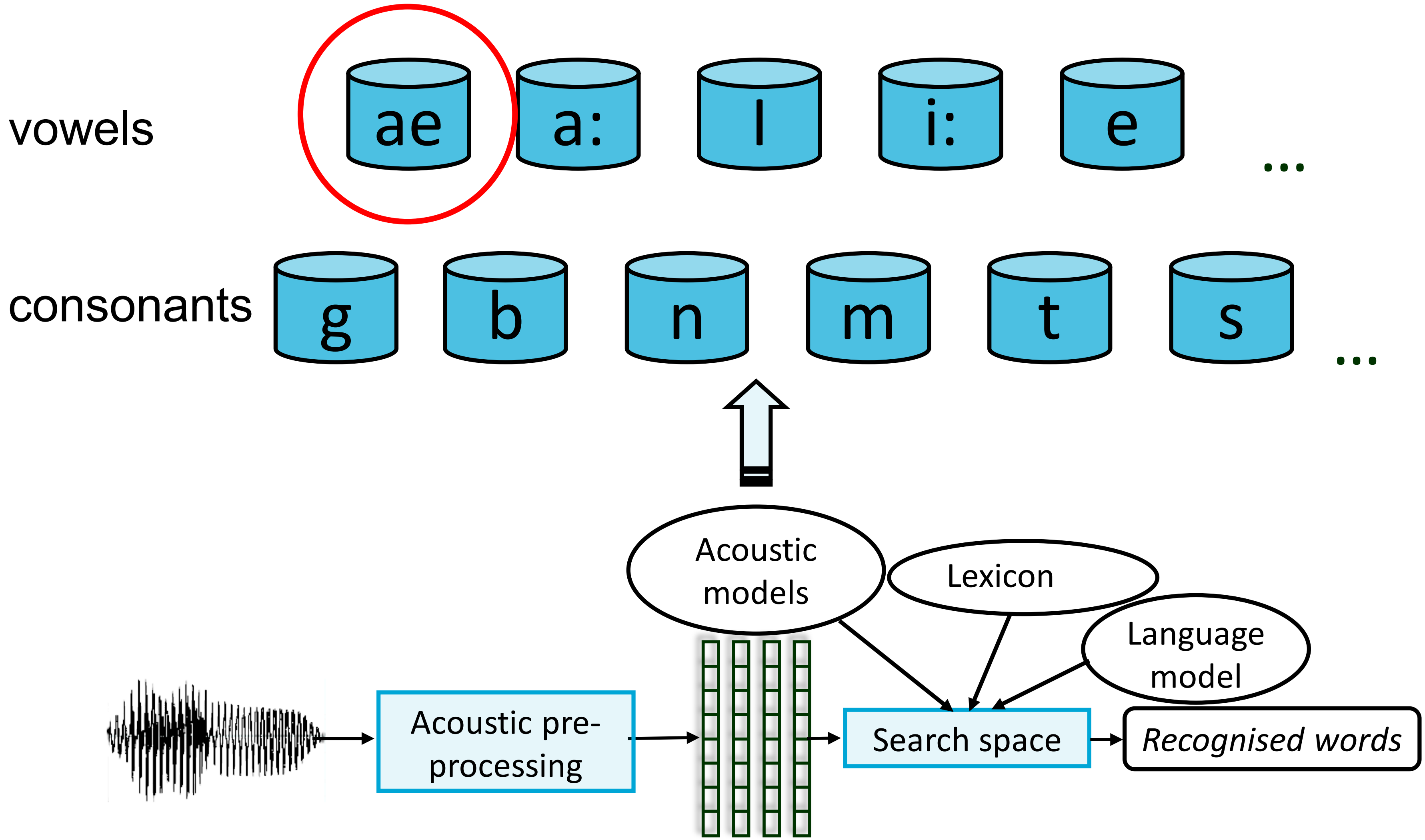
In large vocabulary ASR systems:

- A word consists of multiple phones

For instance: P( X | *tree* )

$$= P(X \mid /t/) \cdot P(X \mid /r/) \cdot P(X \mid /ee/)$$

➔ Derive an estimate of the probability that a particular recognition unit *Phone* generated a particular stretch of speech *X* → P(X|*Phone*)




To compute p(X|phone) → Hidden Markov Models

# Hidden Markov Models

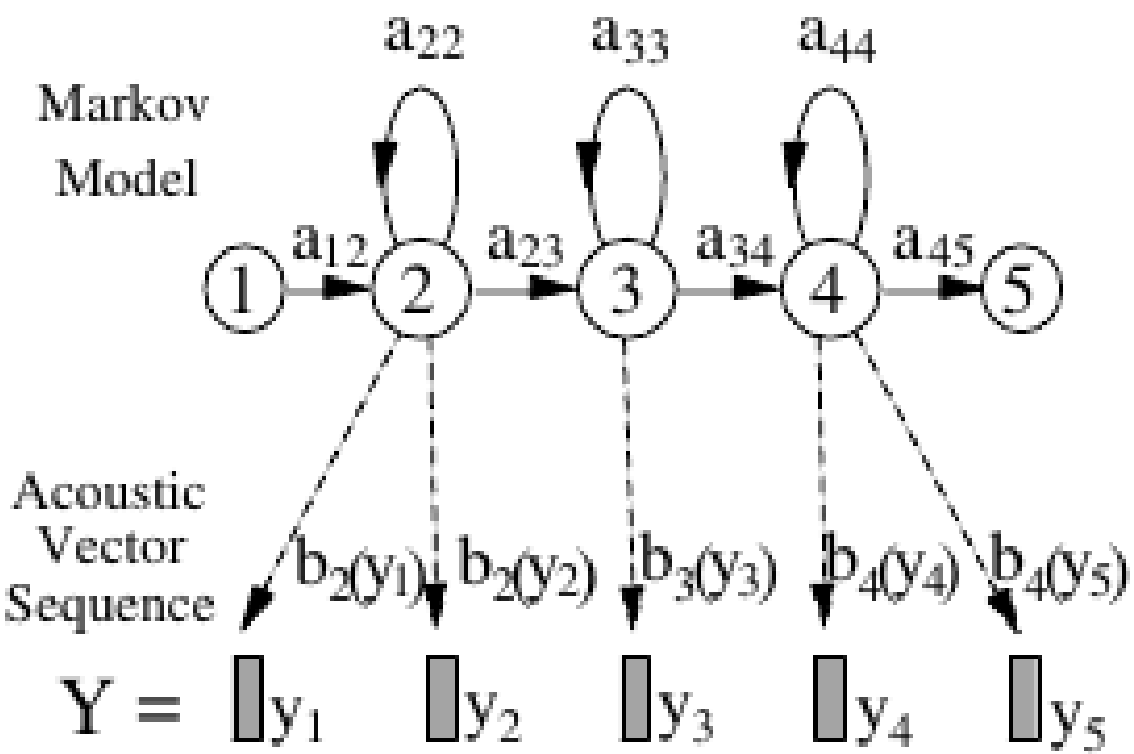
- HMMs can deal with the variability in pronunciations and duration of the speech signal
- Temporal warping of the speech signal is easily done using HMMs, through self-loops
- Assign a probability to an **ambiguous sequence** of **observations**, e.g., a sequence of speech vectors



Slide adapted from Jurafsky & Martin

# Hidden Markov Models (HMMs)

- Statistical model used to calculate p(X|phone)

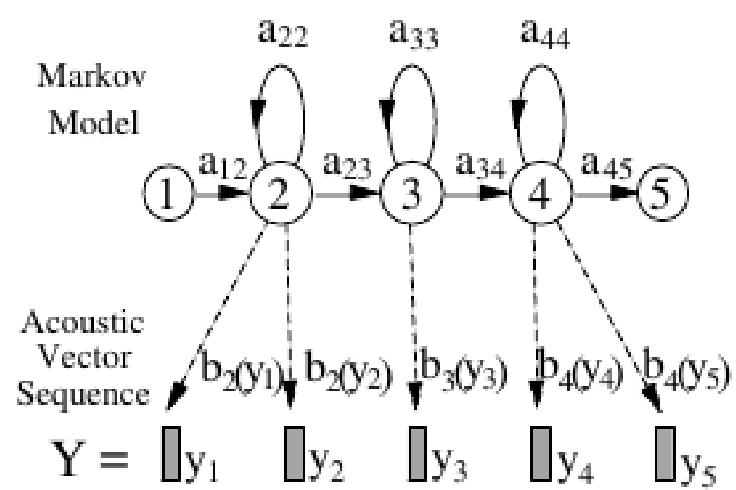


# Hidden Markov Models

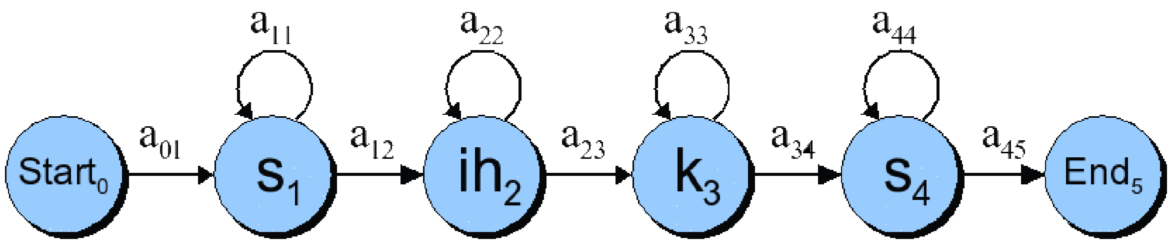
- A set of states: $Q = q_1, q_2 \ldots q_m$; the state at time t is $q_t$
- Start and end state, not associated with observations
- A set of transition probabilities: $A = a_{01} a_{02} \ldots a_{n1} \ldots a_{nn}$
- $a_{ij}$ = the probability of transitioning from state i to state j
- A set of observations: $Y = y_{01} y_{02} \ldots y_{n1} \ldots y_{nn}$
- A set of observation likelihoods (or emission probabilities): $B = b_i(y_t)$ or $b_i(o_t)$

- **First-order Markov assumption:**

Current state only depends on previous state



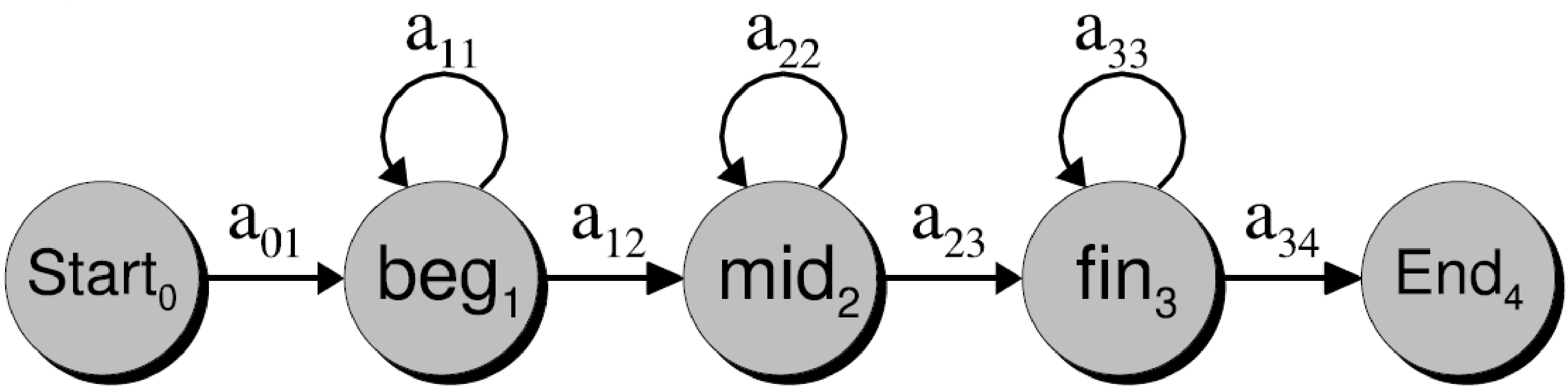
# A very simple HMM




- One state per phone
- Left-to-right
- Typically no state skipping
- Self-loops allow for modelling phone duration

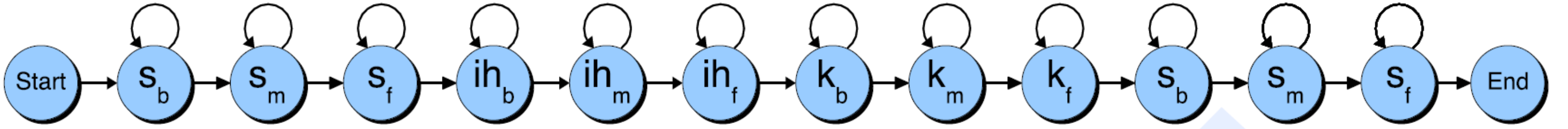
# Multiple states per phone

- Each phone modelled by 3 states + Start + End




TUDelft

# An HMM with 3 states per phone



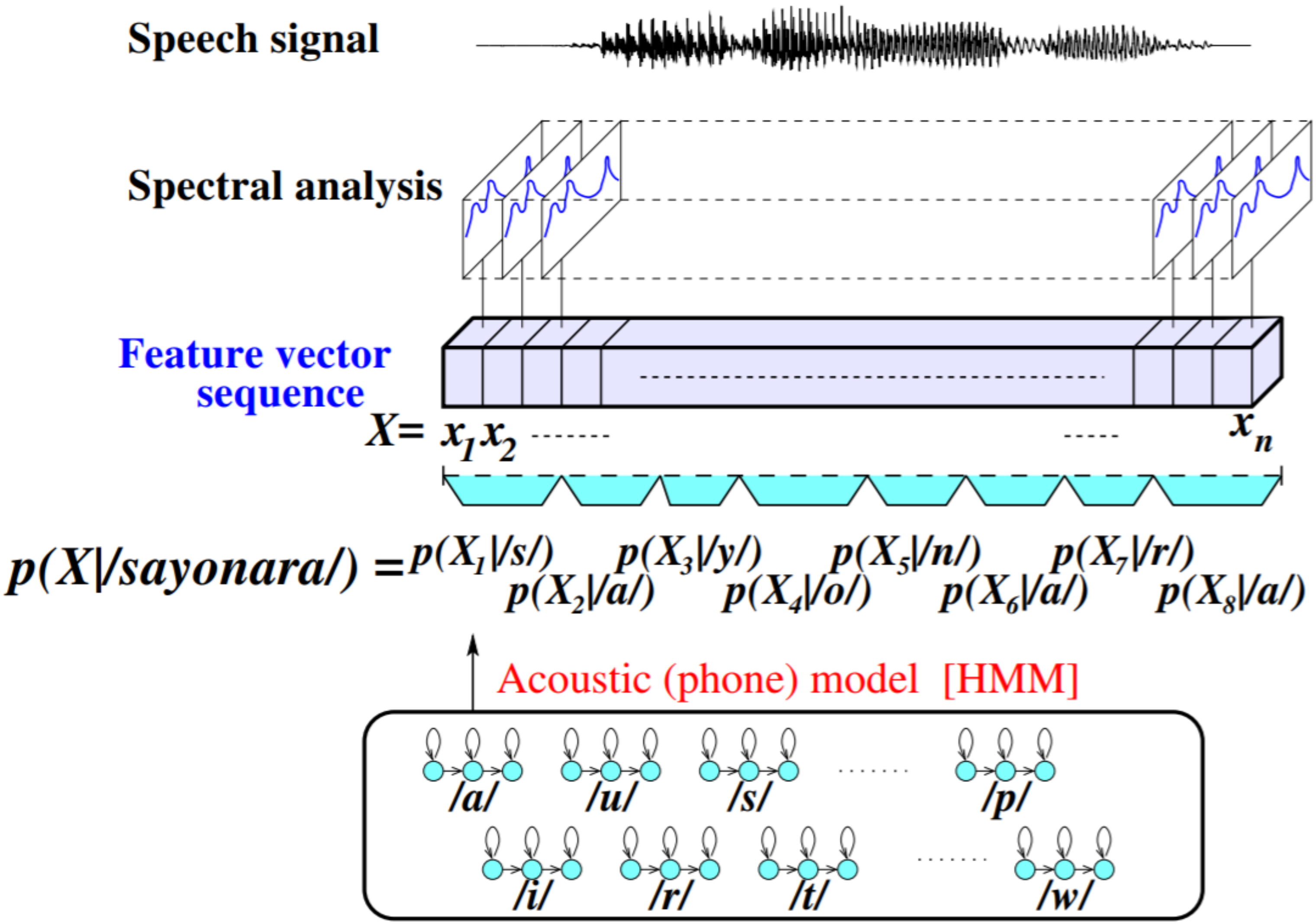
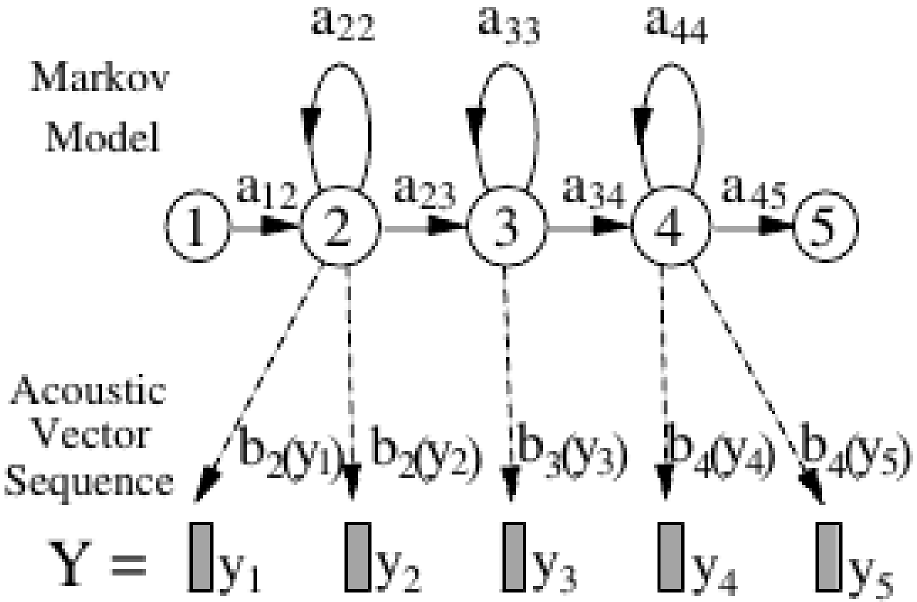


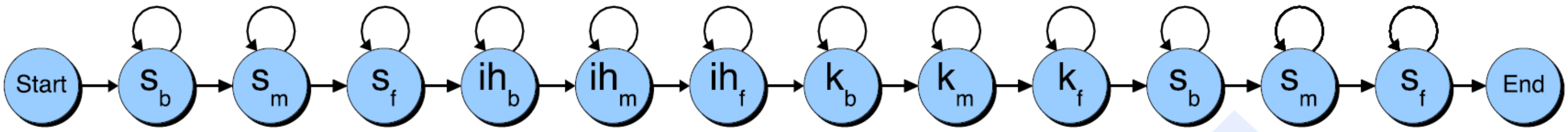

Figure from Shimodaira & Renals, 2017

# Mapping of acoustic features to phones

- The observations (o) (or y below) are the MFCC vectors
- 1 MFCC vector for each frame
- Many MFCC vectors mapped onto one HMM state
- But which MFCC vector is mapped onto which HMM state?



# Which state are we in?




For any given observation of [s ih k s], we could be in multiple states

| $o_1$ | $o_2$ | $o_3$ | $o_4$ | $o_5$ | $o_6$ | $o_7$ | $o_8$ | $o_T$ |
|---|---|---|---|---|---|---|---|---|
| s | s | s | ih | ih | ih | k | k | s |
| s | s | ih | ih | ih | k | s | s | s |
| s | ih | ih | k | k | k | k | k | s |
| ... | | | | | | | | |

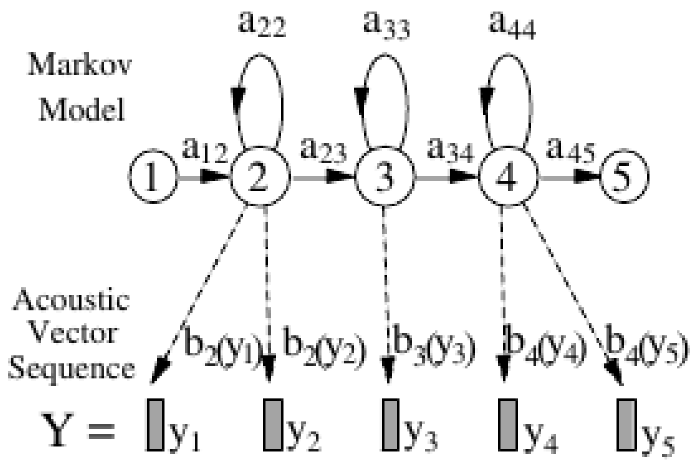
We do not know the mapping, but that is not important

Slide adapted from Jurafsky & Martin

# Training

- Calculate likelihood of a given state *q* generating an observation *o*, i.e., the MFCC feature
  = emission probability $b_j(o)$
  = acoustic likelihood of a frame
  calculated on the basis of a large corpus

- Transition probabilities: from the lexicon




TUDelft

# The emission probability: $b_j(o)$

= likelihood of an observation *o* (MFCC) given a subphone state *q*

- MFCC vectors are real-valued numbers

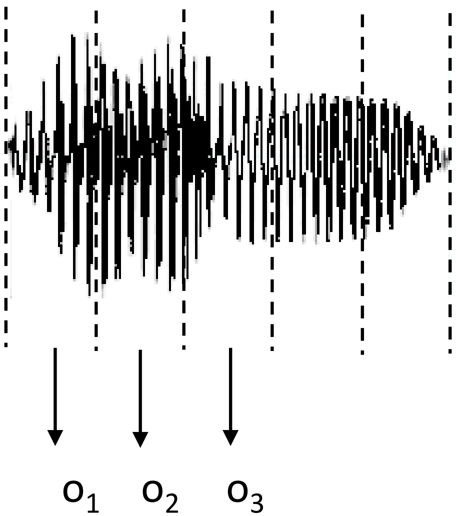
➔ Cannot compute the likelihood of a given state (phone) generating an MFCC vector by counting the number of times each such vector occurs

Can be trained from data using:

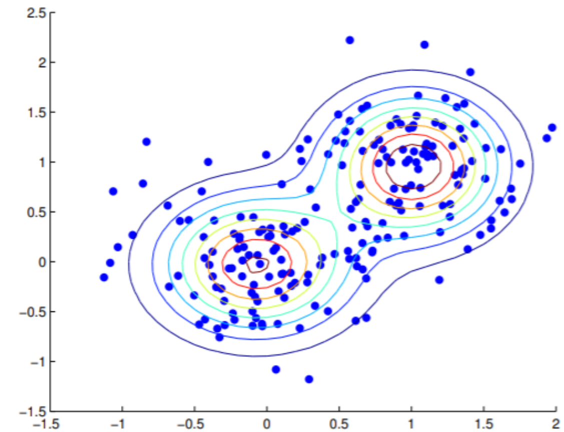
- **Gaussian mixture models**

# How do we train the acoustic models?

We have




We need

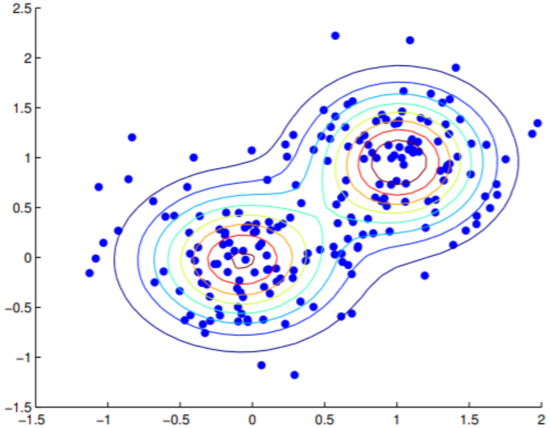



Fitted with a two component GMM using EM

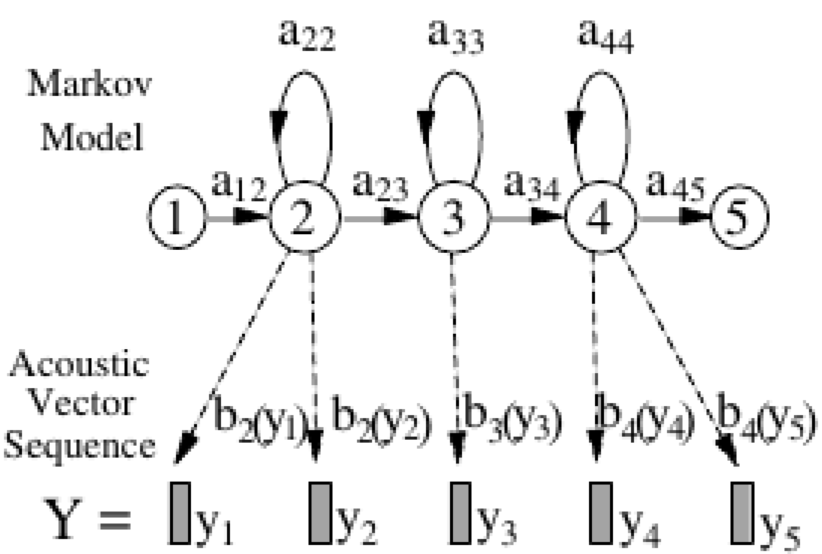
- We need to discover the means and standard deviations of the Gaussians, using HMMs

# Hidden Markov Models

- Remember: the states of an HMM “are” the Gaussian mixture models = B



Fitted with a two component GMM using EM



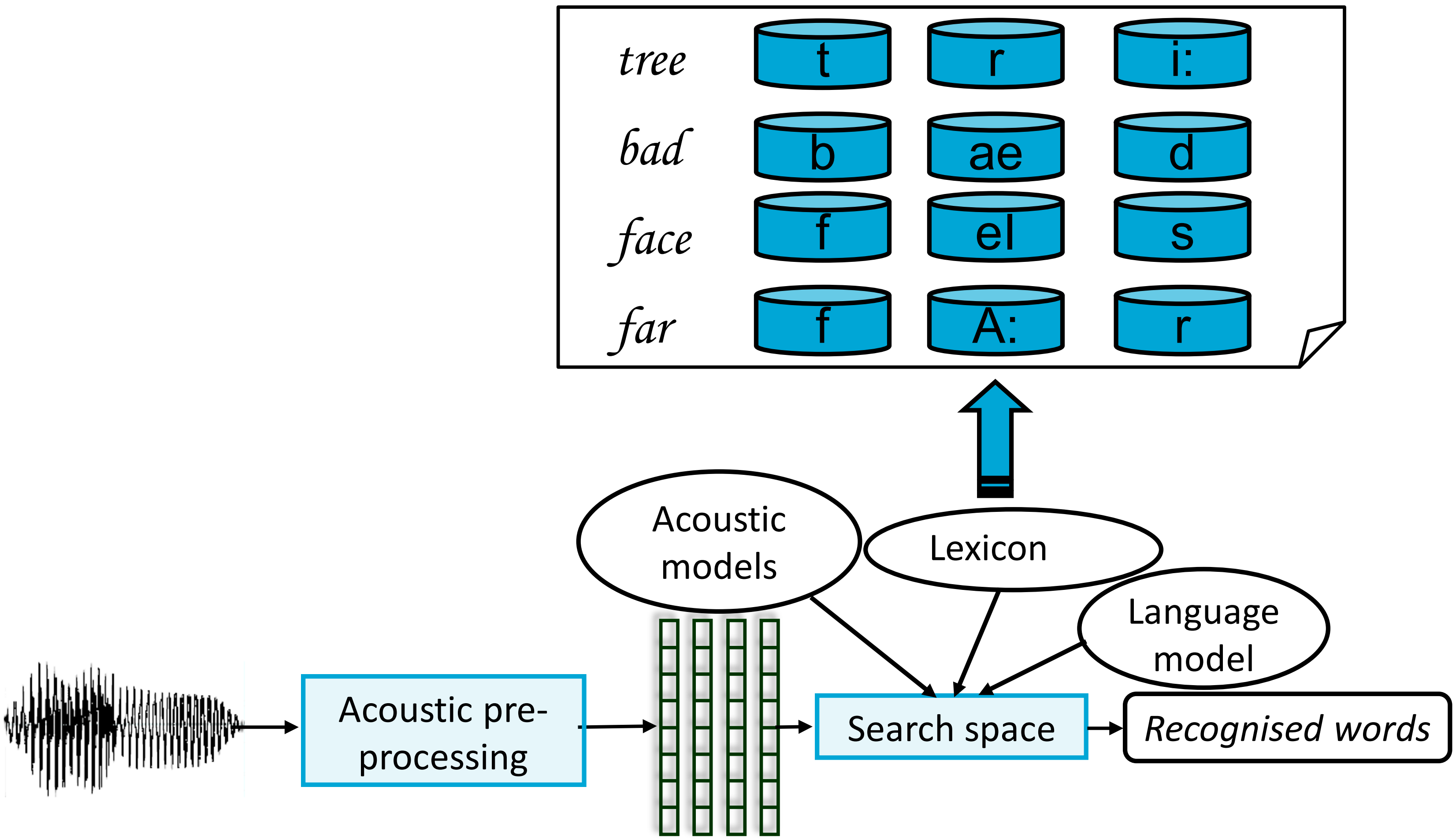
# How to model the constraints

1. Acoustic model: to model the constraints inherent to the speech signal

2. Lexicon: to model the constraints on the order of sounds in a language
   → Determined by the words of a language

3. Language model: to model the constraint inherent to word order in the language

# Lexicon




TUDelft

# Out-of-vocabulary words

- Words in the test corpus that are not included in the lexicon

- OOVs cannot be recognised

- The OOV rate (%) is a lower bound for the word error rate

→ Every OOV word leads to at least one recognition error; average is about 2 errors per OOV word

- Why not add all possible words into the lexicon?

→ Increase in confusability → Increase in the #errors

TU Delft

# How to model the constraints

1. Acoustic model: to model the constraints inherent to the speech signal

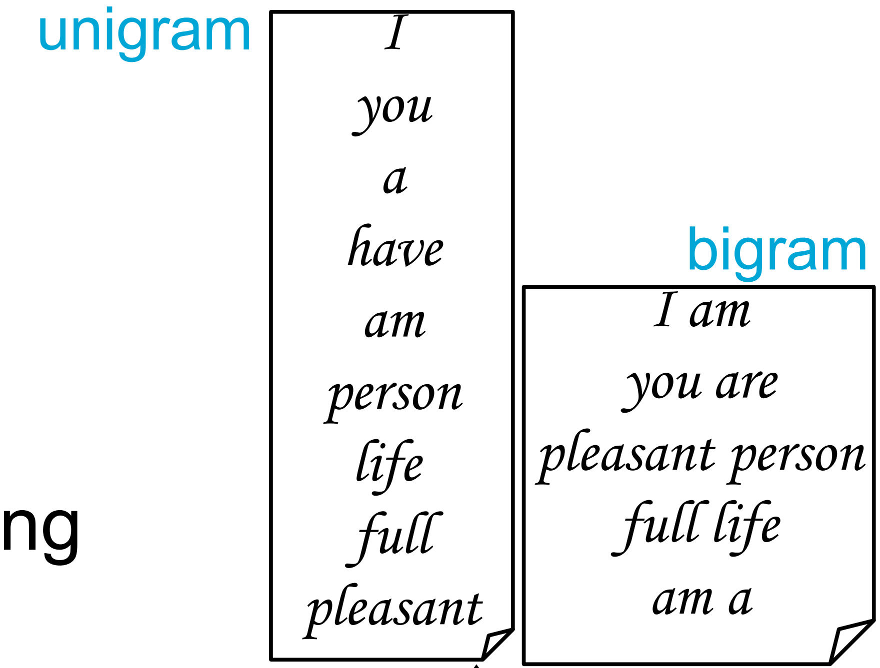
2. Lexicon: to model the constraints on the order of sounds in a language
   → Determined by the words of a language

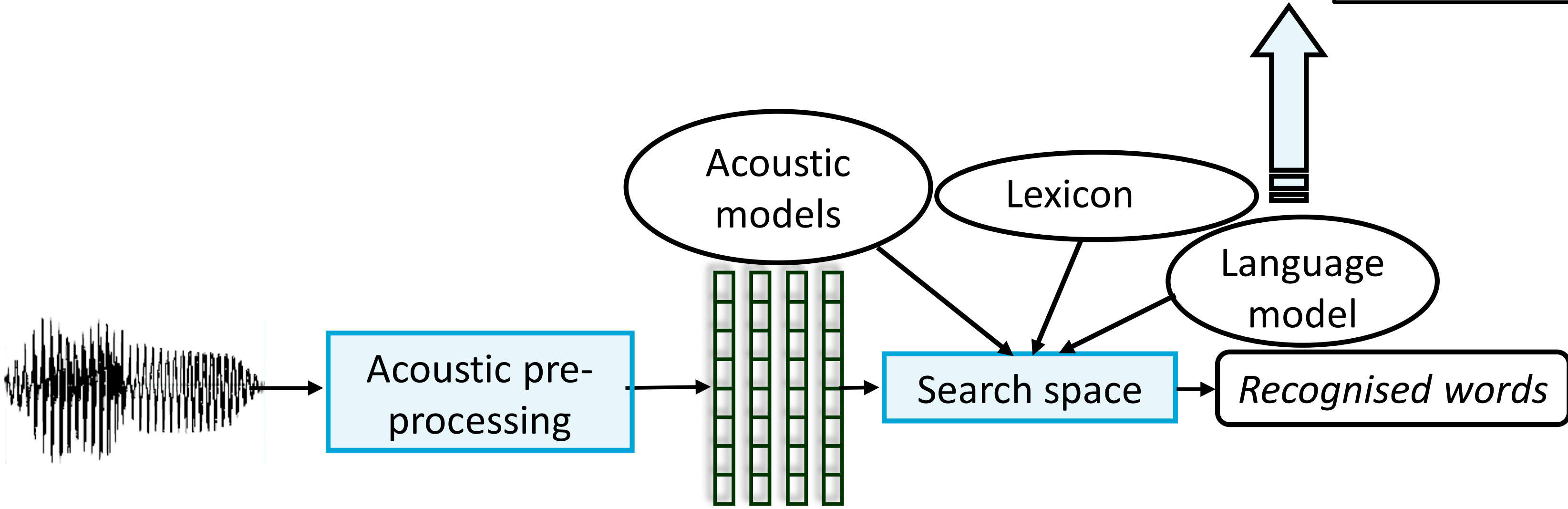
3. Language model: to model the constraint inherent to word order in the language

TUDelft

# Language model

- P(W) = probability of a (sequence of) particular recognition unit(s) occurring



# Training a Language Model

- Choose a language source
- Choose a training set
- Determine the vocabulary
- Estimate the necessary probabilities:
  P(W) = Raw count of W/Total number of running words
- Typically used LMs are 4-, 5-, N-grams

TUDelft

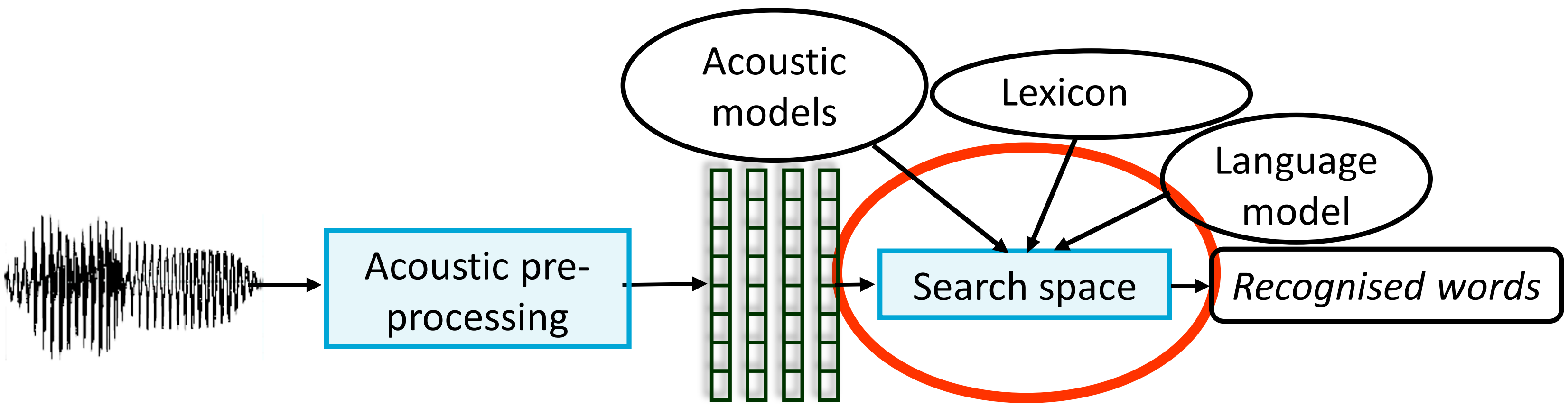
# Decoding

# How to *search* for the most optimal answer: Decoding

What is the **most likely sentence** out of all sentences in the language *L* given some acoustic input X?

= argmax P(W|X) = ( P(X|W) • P(W) )

Output: rank-ordered *N*-best list of most likely word sequences



# Decoding

- Task: simultaneously segmenting the utterance into words and identifying each of these words
- Often done using the Viterbi algorithm

Example: What are the words in this sequence of phones?

[ay d ih s hh er d s ah m th ih ng ax b aw m uh v ih ng r ih s en l ih]

(From Jurafsky and Martin, second edition)

*Answer: I just heard something about moving recently*

- Why is it so hard to segment the speech and identify the words?

# Evaluation and performance

# Evaluation

- On *unseen* data (to check generalisability of the ASR system)
- Dynamic programming to align the ASR output with a reference transcription
- Three types of error: insertion, deletion, substitution
- Word error rate (WER) takes all three types of error into account

# Evaluation

Spoken:

„and that was rather interesting for us as well“

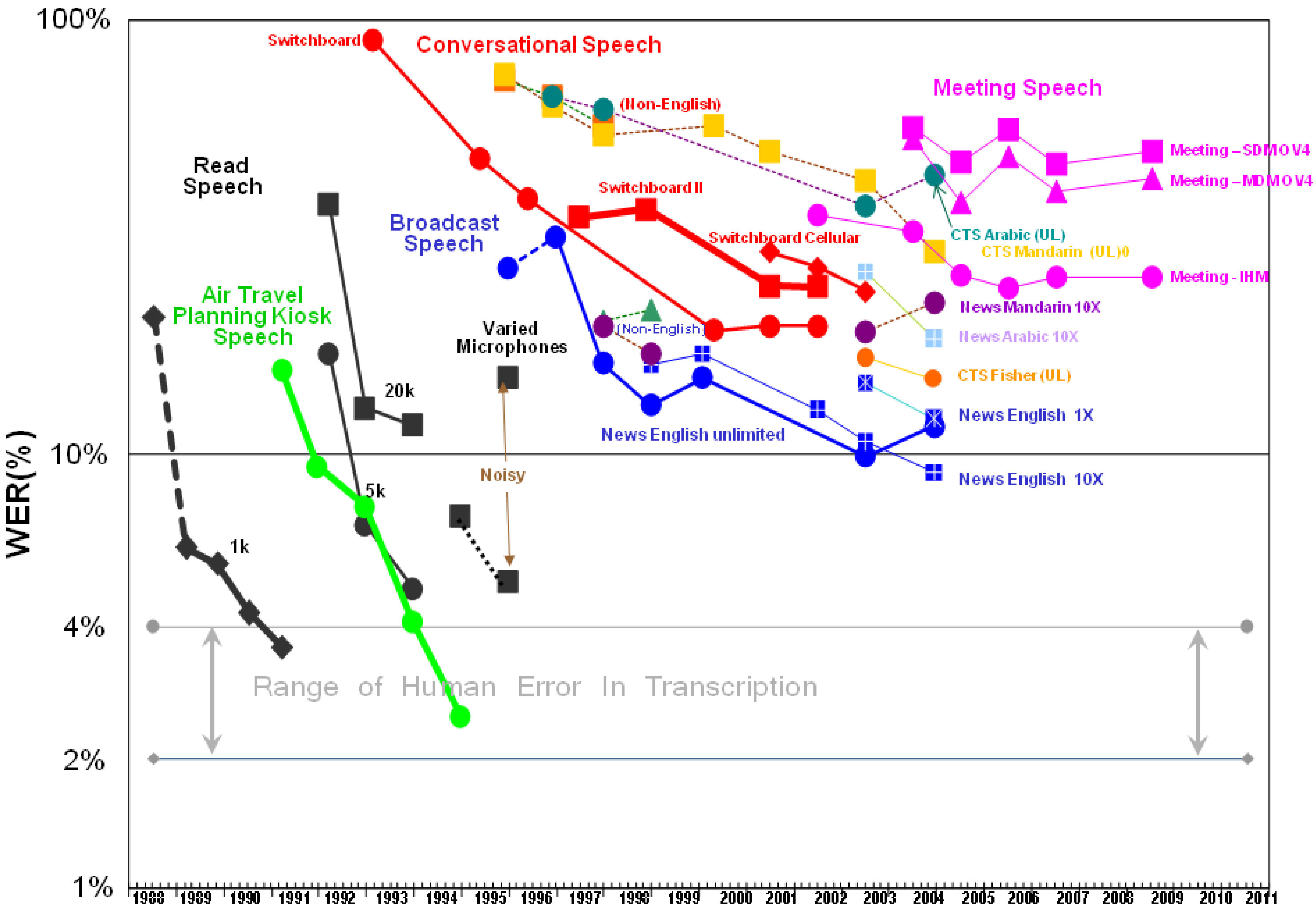
Recognized:

„and that a was father interesting for as well“

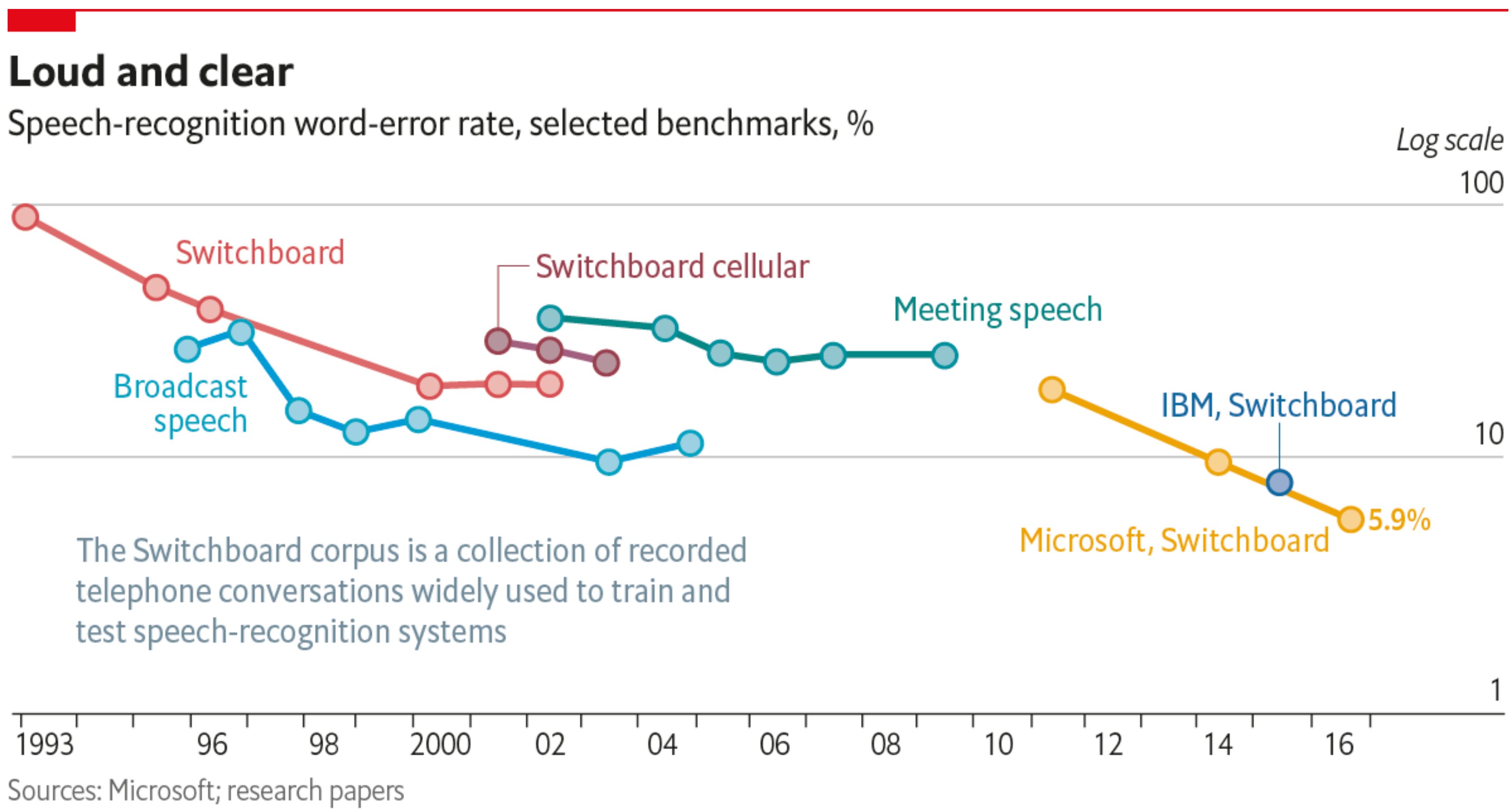
substitution insertion deletion

$$\text{WER} = 100\% \cdot \frac{1 \text{ deletion} + 1 \text{ insertion} + 1 \text{ substitution}}{9 \text{ spoken words}} = 33.3\%$$

# NIST STT Benchmark Test History – May. '09

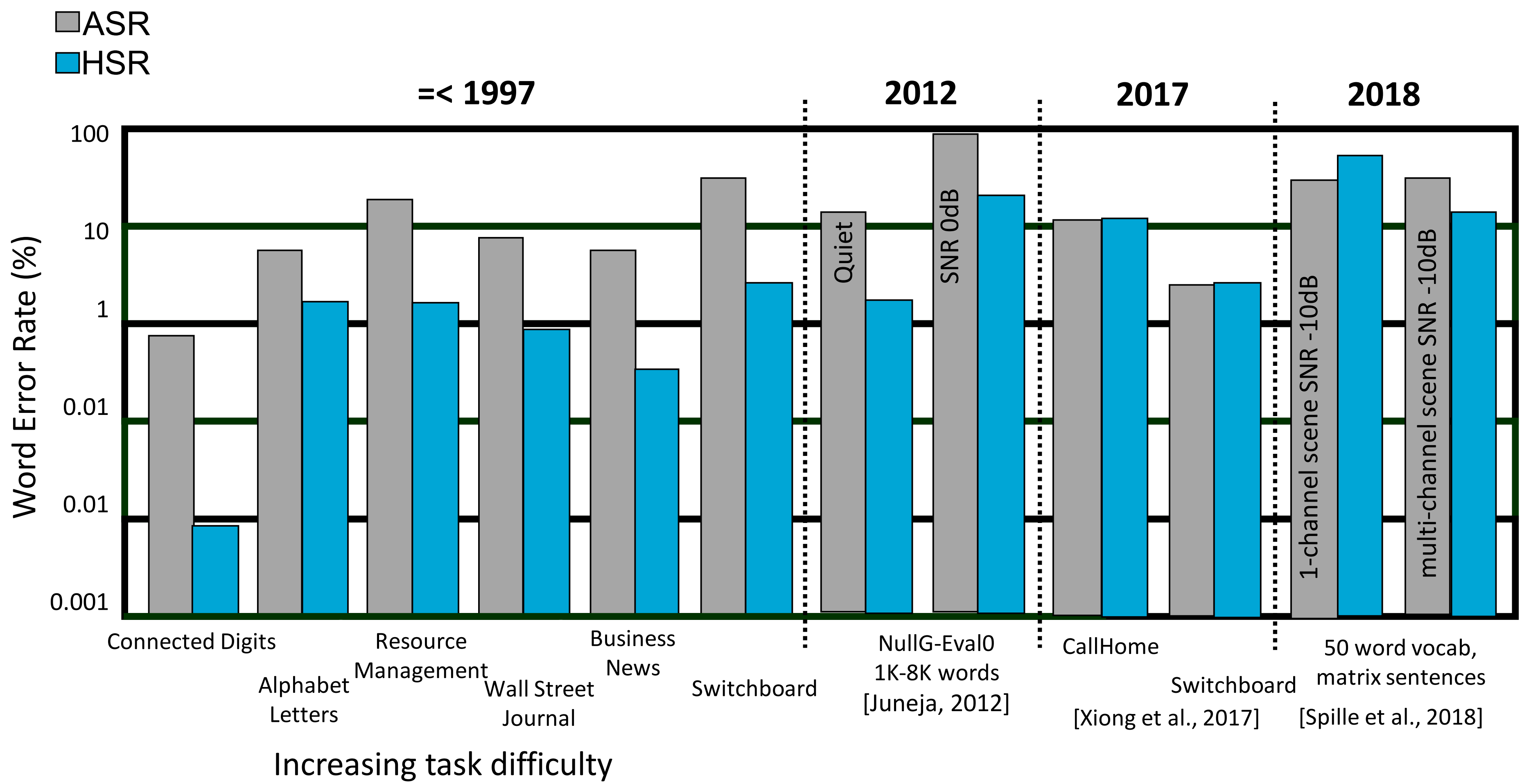



Slide from: https://www.nist.gov/itl/iad/mig/rich-transcription-evaluation




Slide from: https://www.economist.com/technology-quarterly/2017-05-01/language

# Human vs. machine word recognition performance




Data: =< 1997: Lippman, 1997.
Figure extended from: Moore, 2003.

# Limitations of an ASR

- Can you name some?

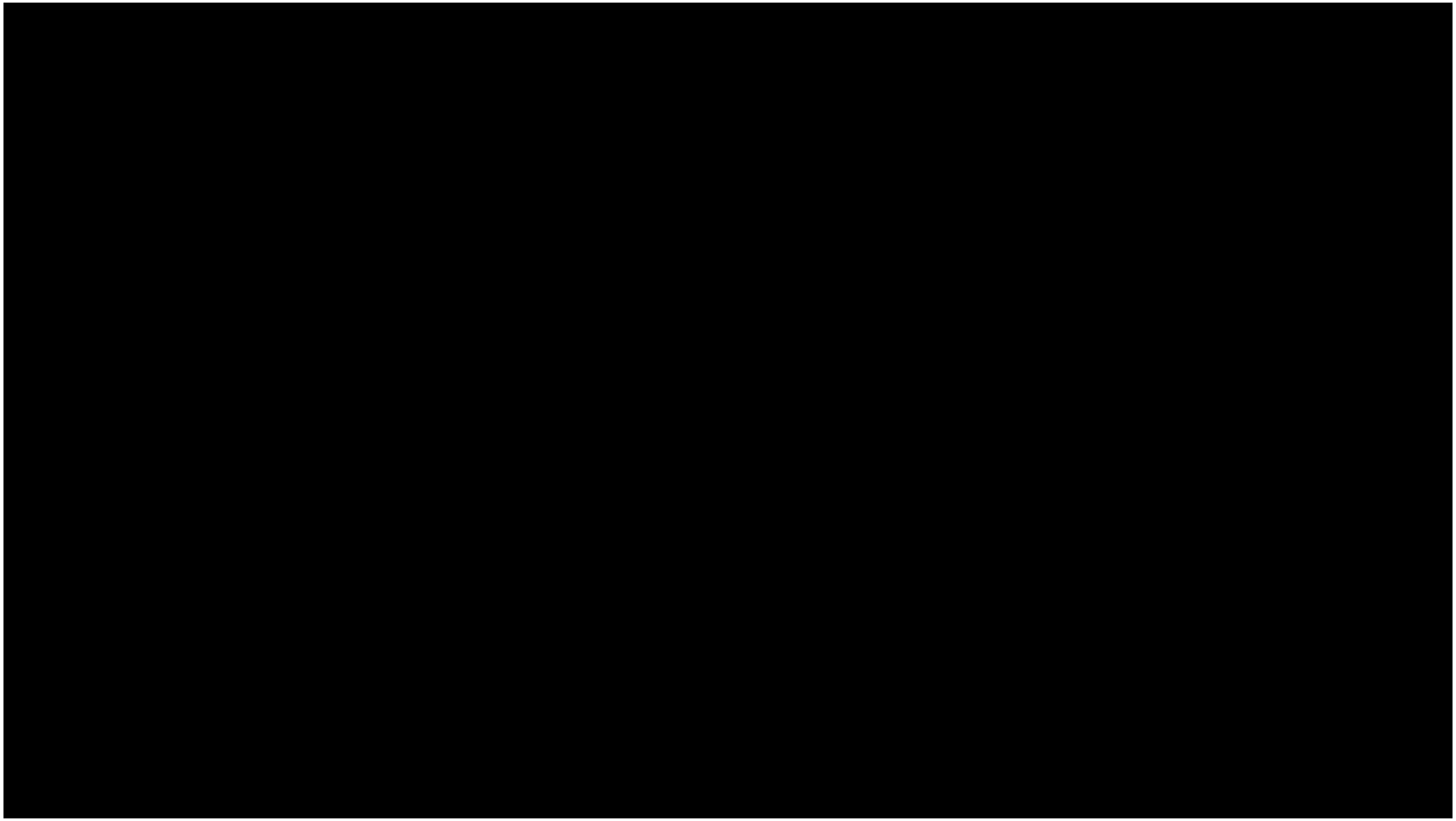
# Limiting factors of ASR

- Continuous signal
- Size of the task:
  - Size of the lexicon
  - Perplexity of the lexicon
- Acoustic environment:
  - Background noise
  - Competing speakers/Overlapping speech
  - Channel conditions (microphone, phone line, room acoustics)
- Speaking style:
  - Isolated words vs. continuous speech
  - Planned speech vs. spontaneous conversation (reductions)
- Speaker:
  - Accents
  - Speaker noises
  - Speaking rate
  - Emotional state
  - Gender
  - Size

# Summary

- ASR = finding the most likely sequence of words given the acoustic signal
- 3 information sources: acoustic models, language models, lexicon → model the constraints of the search space
- Segmentation of the speech signal *follows* from speech recognition
- ASR systems require lots of annotated data, task-dependent

# Limitations of ASR – watch at your own leisure



https://www.youtube.com/watch?v=BOUTfUmI8vs