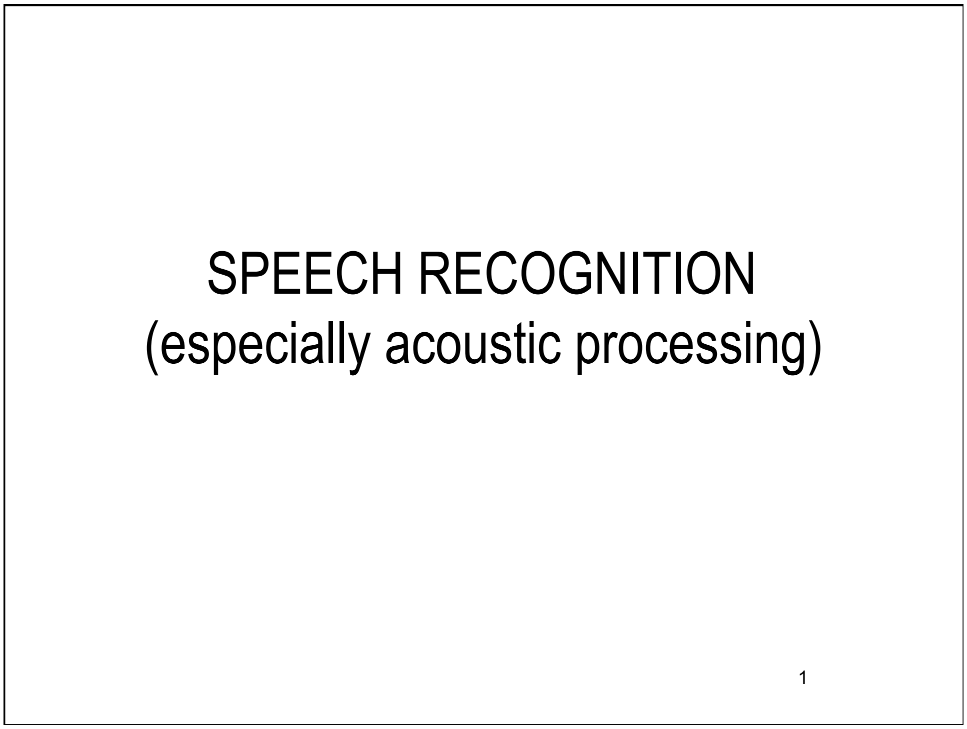
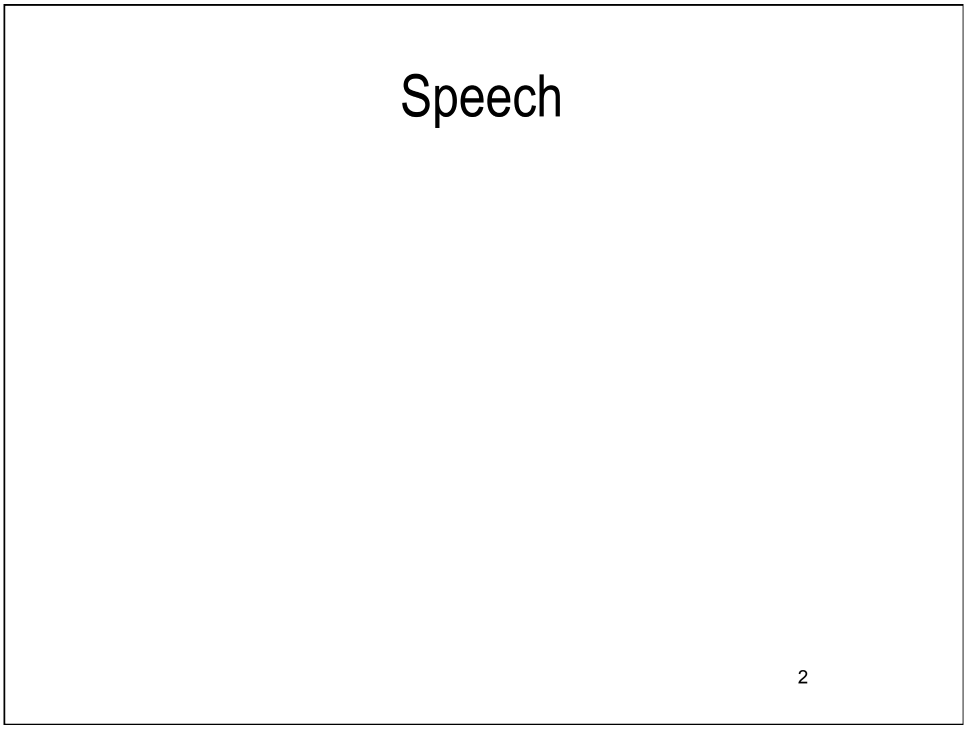
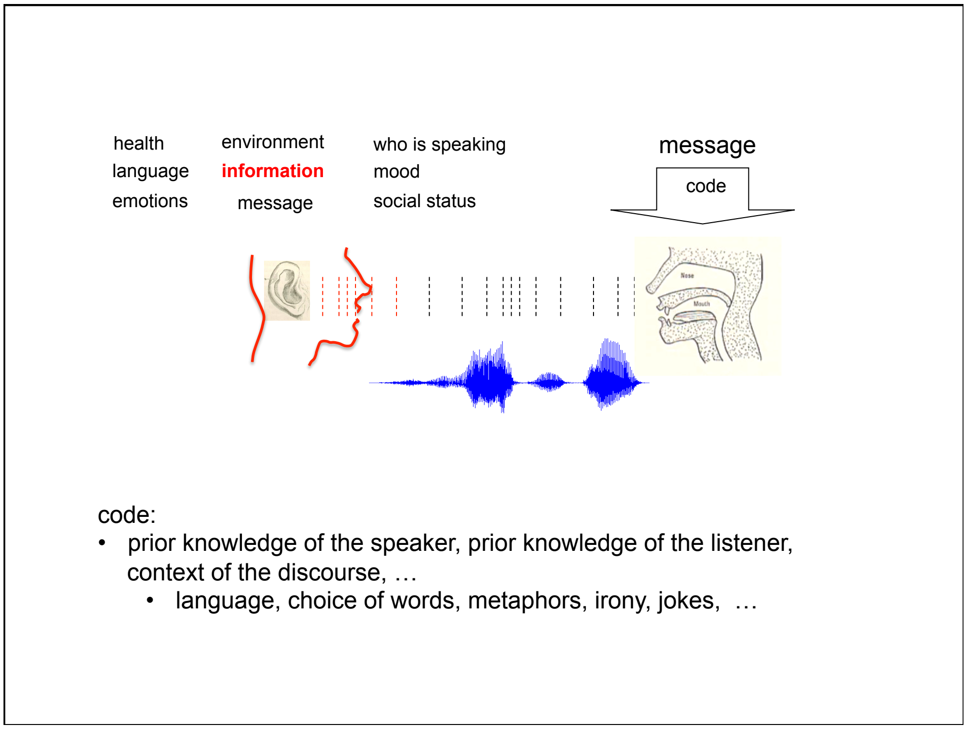


SPEECH RECOGNITION (especially acoustic processing)



Speech



health
language
emotions

environment
information
message

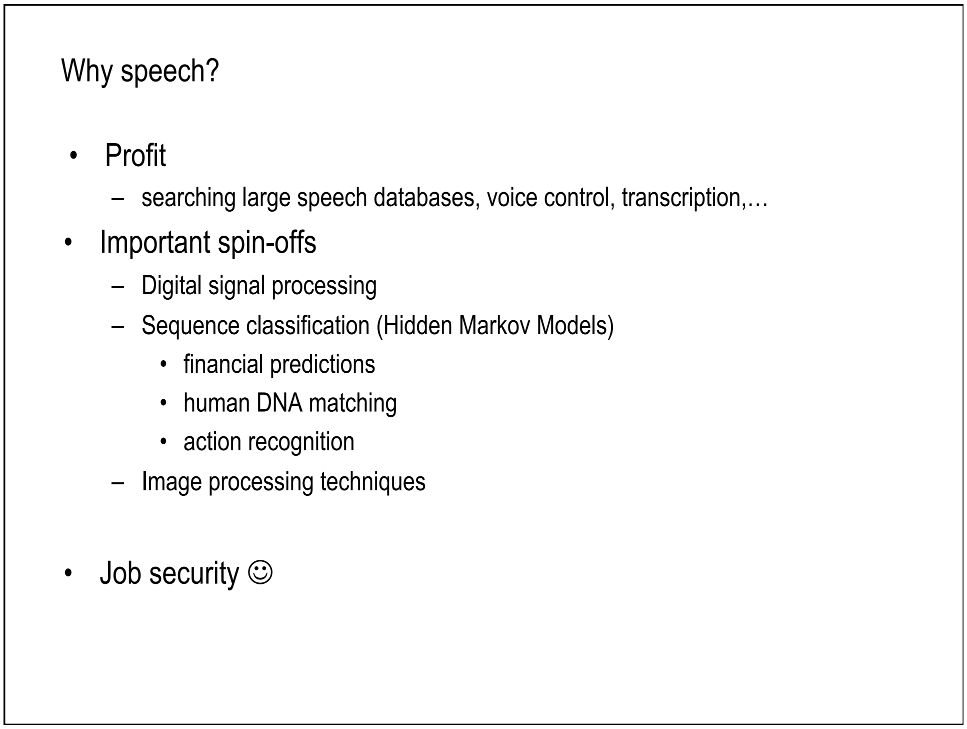
who is speaking
mood
social status

message

code

code:

- prior knowledge of the speaker, prior knowledge of the listener, context of the discourse, …
    - language, choice of words, metaphors, irony, jokes, …



Why speech?

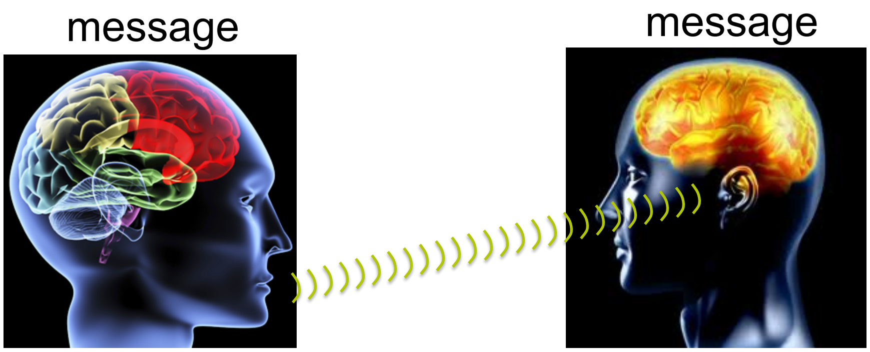
- Profit
    - searching large speech databases, voice control, transcription,…
- Important spin-offs
    - Digital signal processing
    - Sequence classification (Hidden Markov Models)
        - financial predictions
        - human DNA matching
        - action recognition
    - Image processing techniques
- Job security ☺

Because it is there!

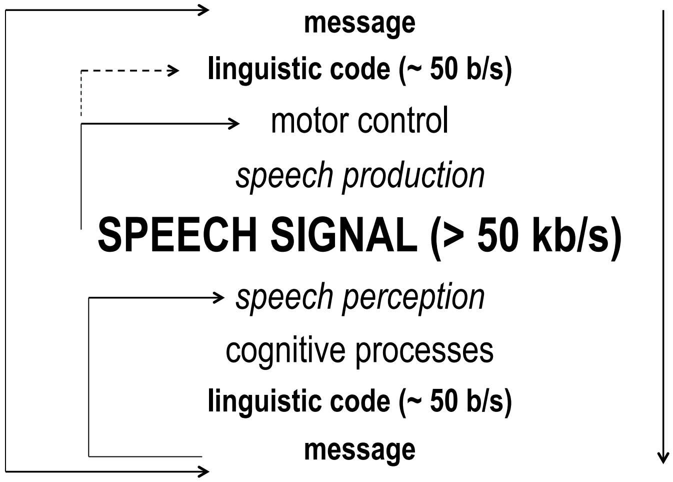
Spoken language is one of the most amazing accomplishments of human race.

Problems faced in machine recognition of speech reveal basic limitations of all information technology !

Spoken Language



Human Speech Communication



message

linguistic code (~ 50 b/s)

motor control

speech production

SPEECH SIGNAL (> 50 kb/s)

speech perception

cognitive processes

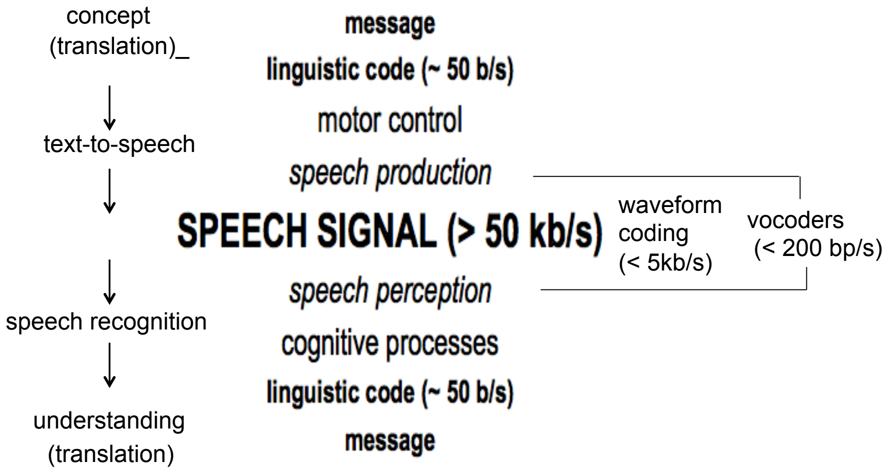
linguistic code (~ 50 b/s)

message

We speak to be heard in order to be understood
Roman Jakobson

Speech Technologies

Speech Engineering



concept
(translation)_

text-to-speech

speech recognition

understanding
(translation)

message

linguistic code (~ 50 b/s)

motor control

speech production

SPEECH SIGNAL (> 50 kb/s)

speech perception

cognitive processes

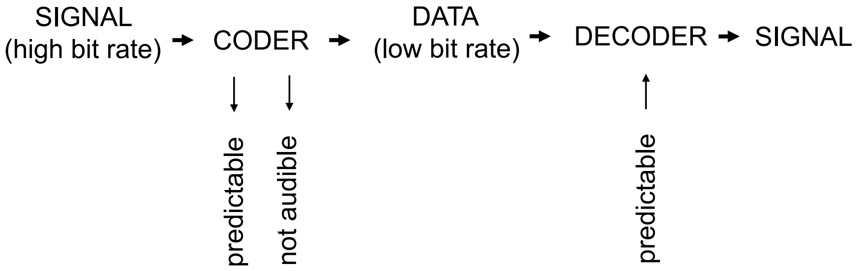
linguistic code (~ 50 b/s)

message

waveform
coding
(< 5kb/s)

vocoders
(< 200 bp/s)

CODING



SIGNAL
(high bit rate) → CODER → DATA
(low bit rate) → DECODER → SIGNAL

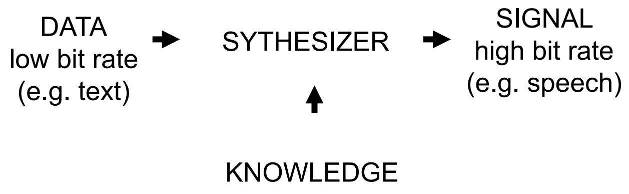
predictable

not audible

predictable

Predictability: from constraints of **speech production** system
Audibility: from constraints of **speech perception** system

TEXT-TO-SPEECH SYNTHESIS



DATA
low bit rate
(e.g. text)

SYTHESIZER

SIGNAL
high bit rate
(e.g. speech)

KNOWLEDGE

KNOWLEDGE

- semantics
- pronunciation
- prosody and intonation
- constraints on speech production (coarticulation)
- constraints on speech perception (what will ear tolerate)
- signal generation (signal processing)

Recognition

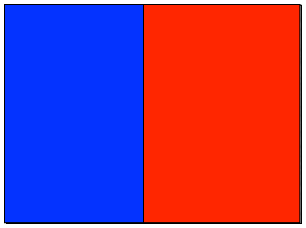
RECOGNITION

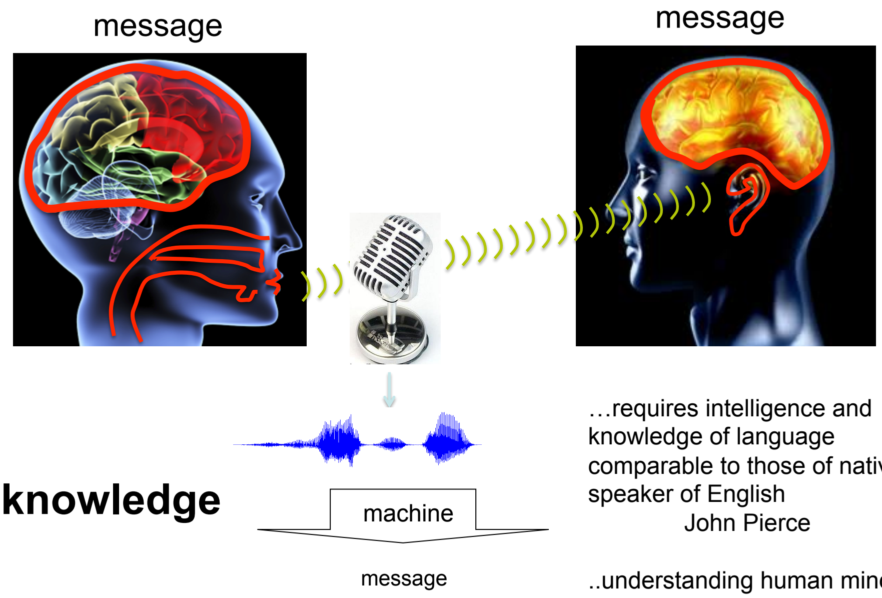
SIGNAL
high bit rate
(e.g. speech) → RECOGNIZER → INFORMATION
low bit rate
(e.g. text, speaker,language, health)

↑

KNOWLEDGE

- Decrease in bit rate = reduction of information content (reduction of signal entropy)
- Second law of thermo-mechanics
  - on its own, entropy cannot decrease





knowledge

…requires intelligence and knowledge of language comparable to those of native speaker of English

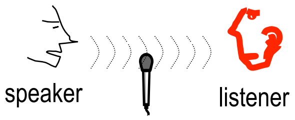
John Pierce

..understanding human mind?

Human mind

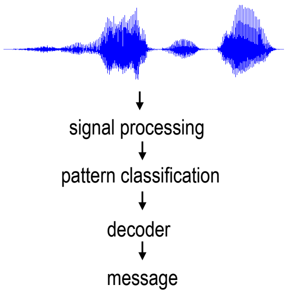
"Like a sword that cuts, but cannot cut itself;
Like an eye that sees, but cannot see itself."
~ *Text from the Zenrin Kushu* ~
(compiled 1429 – 1504)

Nevertheless, nobody can stop us from trying ☺



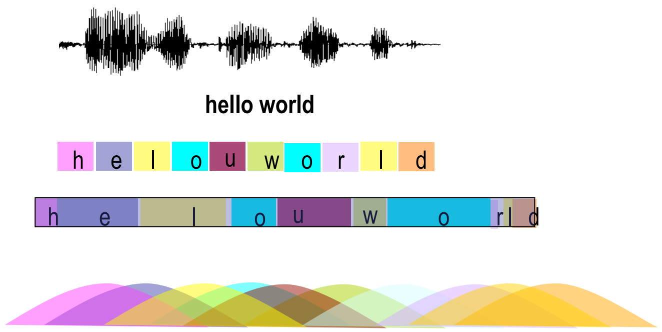
A message is sent by an addresser to an addressee. For this to occur, the addresser and addressee must use a common code, a physical channel, or contact, and the same frame of reference, or context.

We speak in order to be heard in order to be understood



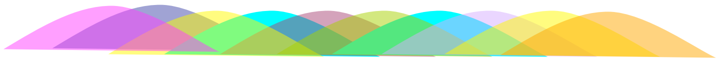
signal processing
↓
pattern classification
↓
decoder
↓
message

Speech recognition
…a problem of maximum likelihood decoding



coarticulation+ talker idiosyncrasies + environmental variability = **a big mess**

Engineering solution



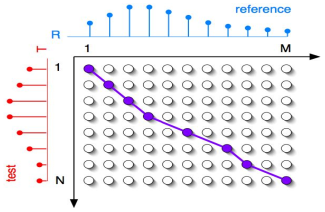
Introduce “context-dependent” phonemes

- class depends on the phoneme identity and on its neighbors

Decision about what has been said is made on **global match** of the incoming data and the data generated by the model, **combined with prior knowledge** (language model)

- allows for local mismatches

Decision is made on the whole speech segment, **subsegment elements (words, syllables, phonemes) obtained as a byproduct of the process**



- find the most likely path aligning the test data and the data generated by the model

Hidden Markov Model

Two dominant sources of variability in speech

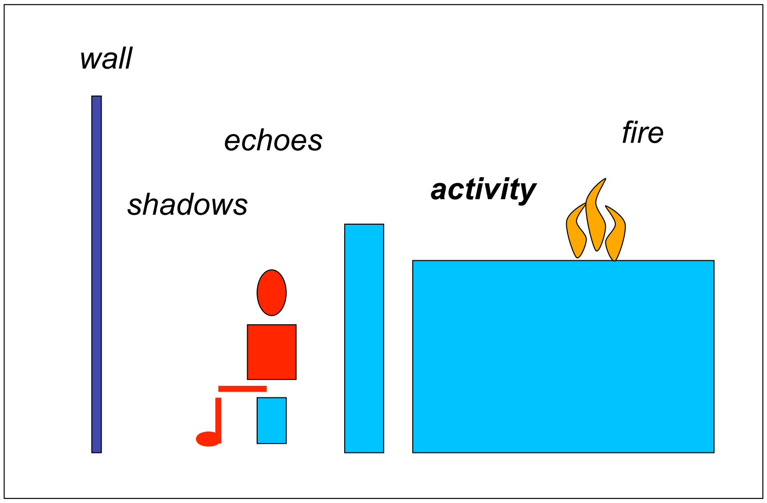
1. different people sound different, communication environment different,… (feature variability)
2. people say the same thing with different speeds (temporal variability)

“Doubly stochastic” process (Hidden Markov Model)

Speech as a sequence of hidden states (phonemes) - recover the sequence

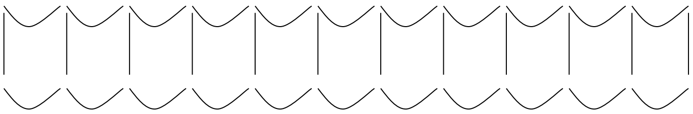
1. never know for sure which data will be generated from a given state
2. never know for sure in which state we are

already old Greeks ……..

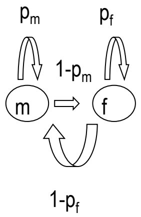


f_0= 195 125 140 120 185 130 145 190 245 155 130 Hz

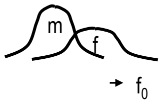
hi hi hi hi hi hi hi hi hi hi hi



know



P(sound|gender)



p_{1m} P(gender)

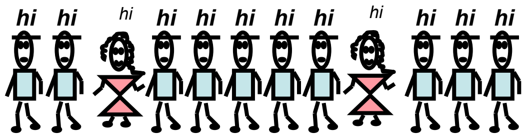
These parameters are typically learned from training data.

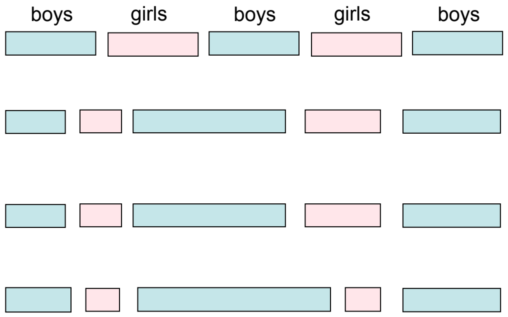
Want to know

where are the boys (or girls) ?

Training of the model

f_0=140 120 190 125 155 130 145 160 245 165 135 150Hz



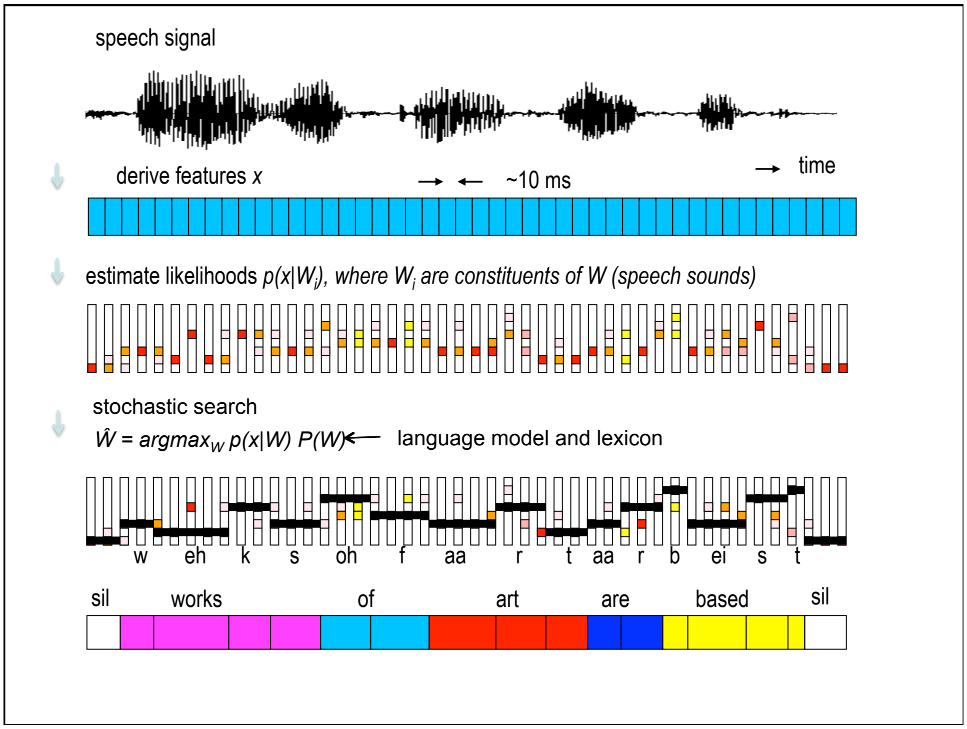


for equally distributes states, compute distributions of parameters for each state

find the best alignment of states given the parameters

compute distributions of parameters for each state

find the best alignment of states given the parameters



Stochastic machine recognition of speech

$$w = \underset{i}{\arg\max}(P(M(w_i) \mid x))$$

How to find w ?

Form of the model $M\,(\,w_i\,)$?

What is the data x ?

How to find w ?

$$w \propto \arg\max_{i} (p(x \mid M(w_i)P(M(w_i)^{\gamma})$$

Form of the model ?

M(w$_i$) – model of the whole
utterance

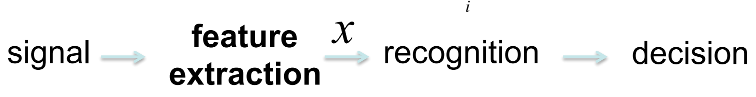
Good: parts of the utterance can be corrupted and the utterance can still be correctly recognized

Bad: low prior probability items in the utterance may be substituted by wrong ones

Ugly: words that are not in the vocabulary will **never** be recognized

What should be the x ?

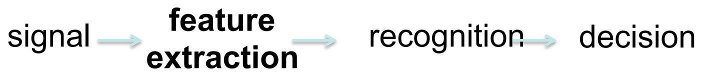
$$w \propto \arg\max_{i} (p(x \mid M(w_i)P(M(w_i)^{\gamma})$$



signal → **feature extraction** $\xrightarrow{x}$ recognition → decision

signal
contains wanted and unwanted variability (information)
may be in a form that is not suitable for the recognition stage

features
what is lost is lost forever
what is kept may cause problems later


signal → **feature extraction** → recognition → decision

signal

- contains wanted and unwanted variability (information)
- may be in a form that is not suitable for the recognition stage

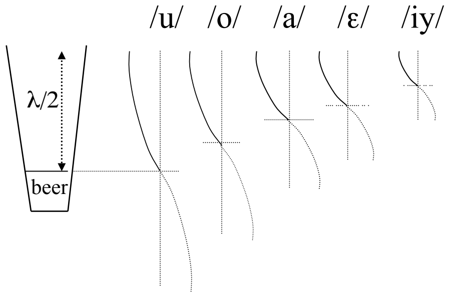
features

- what is lost is lost forever
- what is kept may cause problems later

History

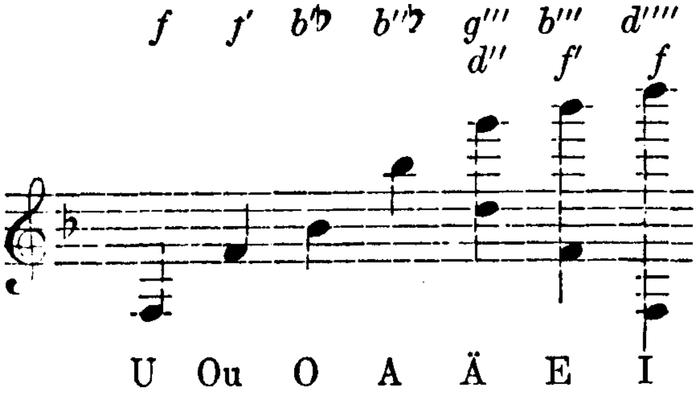
Understand sources of information

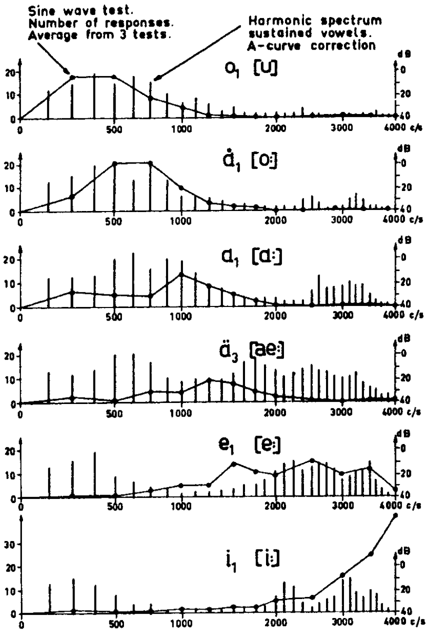
Isaac
Newton



- Extract prime information-bearing elements
  - Helmholtz, Scripture,...

Helmholtz and two-tone vowels

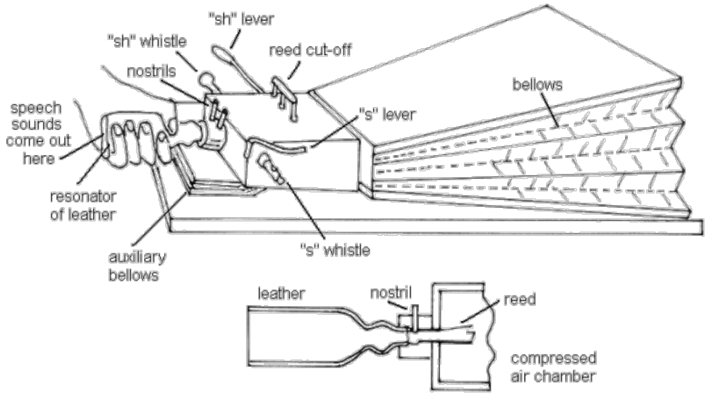




- affiliate vowel with sine wave tone (forced judgement)
- peak of histograms would correspond to resonance frequency of uncoupled front cavity in production of a given vowel
  - Fant 1947

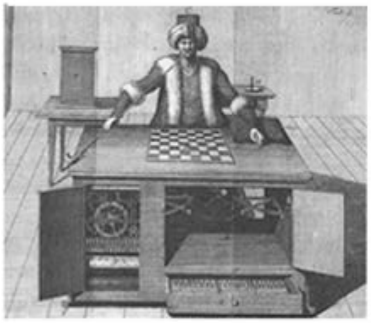
Producing speech

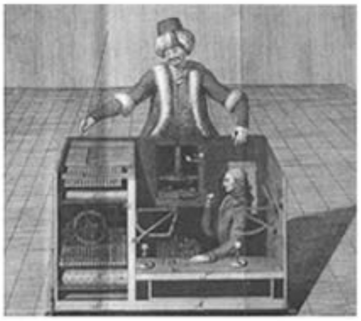
Johann Wolfgang Ritter **von Kempelen** de Pázmánd



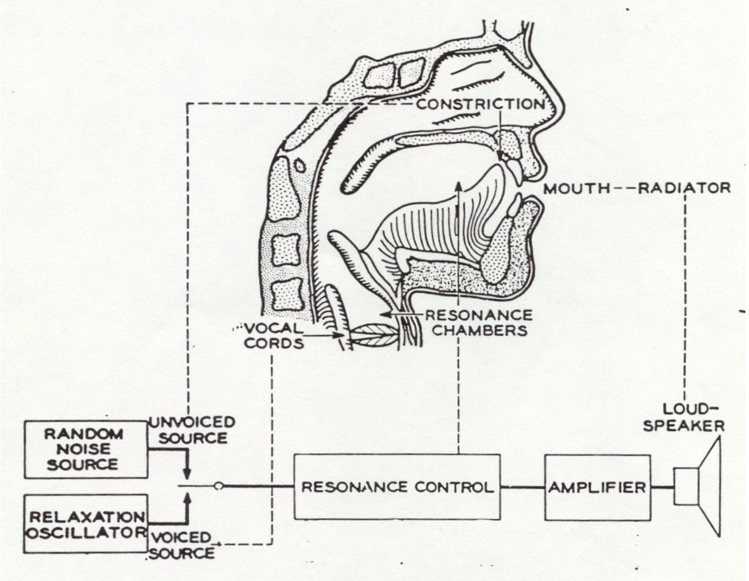
Mechanical Turk

Johann Wolfgang Ritter **von Kempelen** de Pázmánd

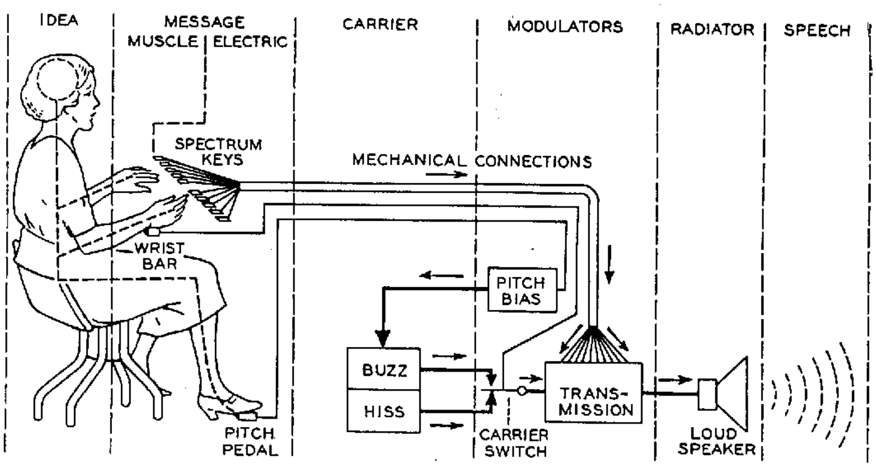




Speech production

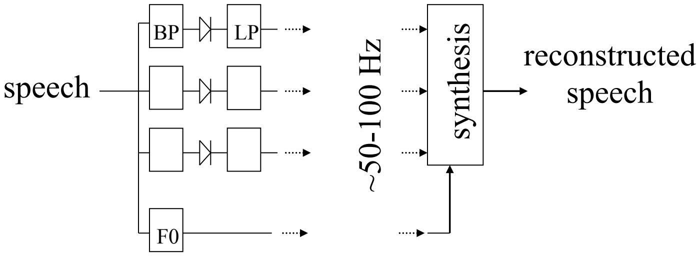


VODER
(Homer Dudley 1940)

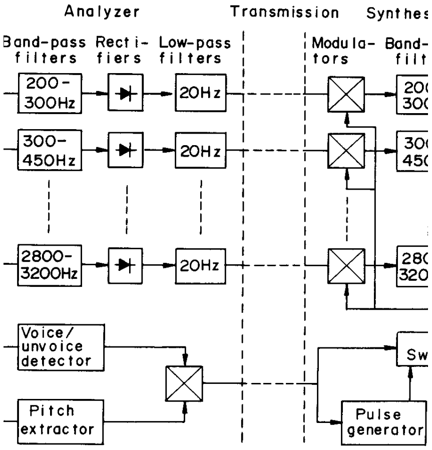


Low bit transmission/storage

- Extract perceptually-relevant components of the signal
- Separate into elements which can be easily described, well quantized, and slowly updated



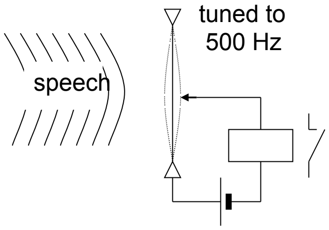
VOCODER
(Homer Dudley 1939)

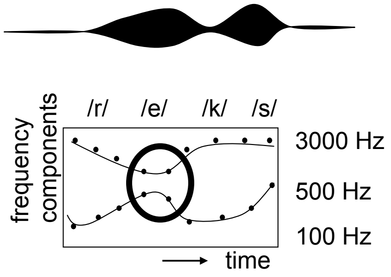


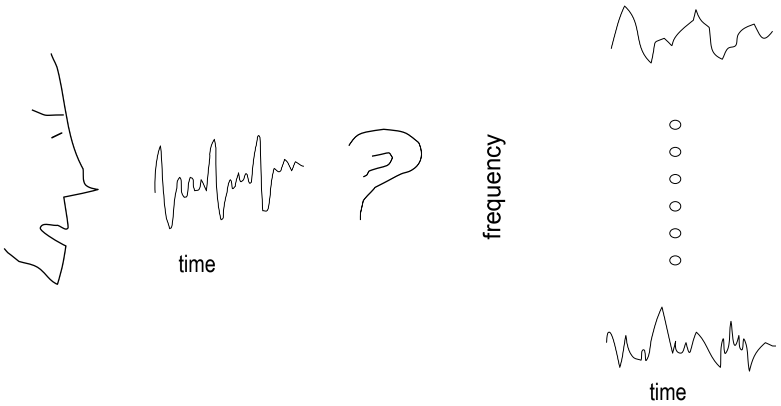
- Predictability (production)
  - speech waveform changes "slowly" (inertia of air mass in vocal tract cavities)
  - spectral envelope changes slowly
    - 20 Hz low-pass
  - voiced speech is periodic
    - pulse generator for excitation
- Hearing properties (perception)
  - spectral resolution of hearing
    - wider band-pass filters at higher frequencies

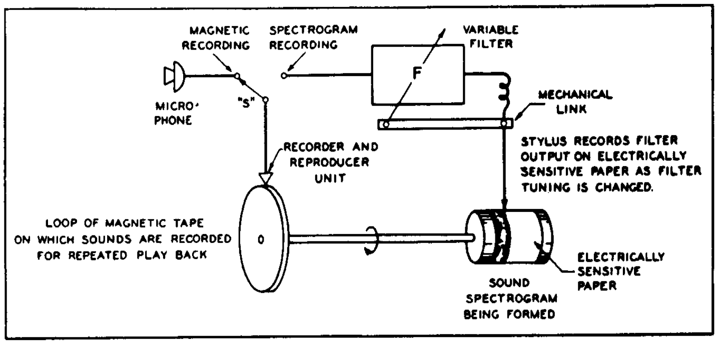
Speech recognition

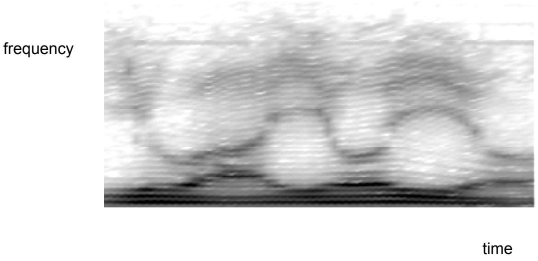
Radio Rex (1917)






time
frequency
time


MAGNETIC RECORDING
SPECTROGRAM RECORDING
VARIABLE FILTER
F
MICRO-PHONE
"S"
MECHANICAL LINK
STYLUS RECORDS FILTER OUTPUT ON ELECTRICALLY SENSITIVE PAPER AS FILTER TUNING IS CHANGED.
RECORDER AND REPRODUCER UNIT
LOOP OF MAGNETIC TAPE ON WHICH SOUNDS ARE RECORDED FOR REPEATED PLAY BACK
0
ELECTRICALLY SENSITIVE PAPER
SOUND SPECTROGRAM BEING FORMED


frequency
time