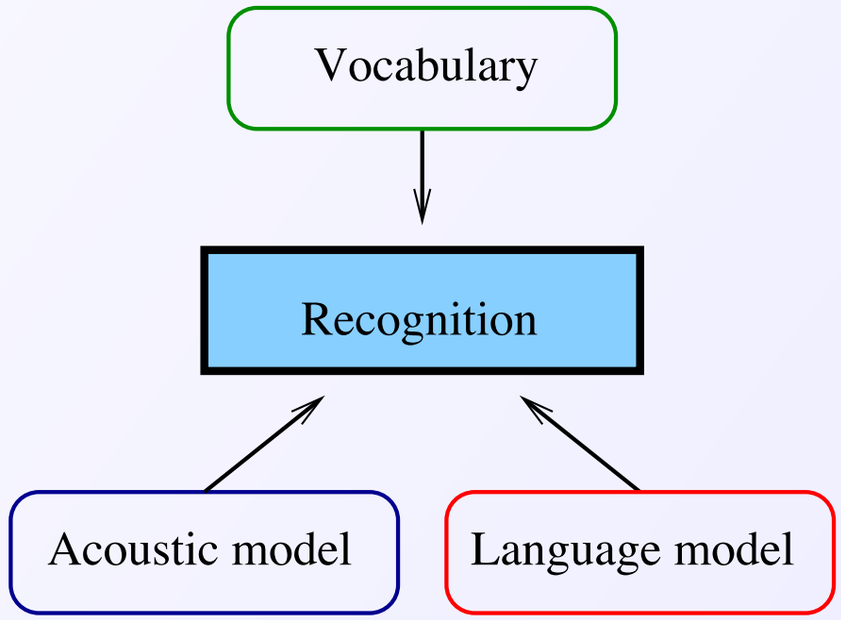
COM4511/COM6511 - Speech Technology

L1
Introduction to Speech Technology



Thomas Hain
t.hain@sheffield.ac.uk
Spring Semester

Module Outline

Lecturers Thomas Hain and Anton Ragni

Duration 6 weeks

Lectures and practical sessions

	Monday	Wednesday
12-13	Broad Lane Block LT 11	-
13-14	Broad Lane Block LT 11	Broad Lane Block LT 11
14-15	-	Broad Lane Block LT 11
17-18	Comp Room F11	

Note

- ☑ 10 lecture sessions / 5 lab sessions
- ☑ Updated lecutre program

Module - Fundamentals

Lectures

20 lectures

Practical work (python)

1 Assignment - choose 5 tasks - from 6+
Effort approx 1 day / task.

Specific reports on each.

Assessment

70% Written Exam, 30% Practical

Material

1. Handout copies
2. Practical lab-sheets
3. Software and data

Lecture plan

Week	Lecture	Topic	Who
1	L1	Intro & Speech Recognition	TH
	L2	Front-ends	TH
	L3	Far-field & Enhancement & Noise	TH
	L4	From distances to distributions	TH

Week	Lecture	Topic	Who
2	L5	Neural networks	AR
	L6	From points to sequences	AR
	L7	HMMs	TH
	L8	Acoustic models	TH

Lecture plan (2)

Week	Lecture	Topic	Who
3	L9	Acoustic models: advanced NN forms	AR
	L10	Language models	AR
	L11	Language models: advanced NN forms	AR
	L12	Search	TH

Week	Lecture	Topic	Who
4	L13	Integrating with others: lattices, confidences	AR
	L14	Un-, semi- and lightly supervised training	AR
	L15	Diarisation	TH
	L16	Adaptation	TH

Lecture plan (3)

Week	Lecture	Topic	Who
5	L17	Speaker ID	TH
	L18	Speech Synthesis	AR
	L19	Speech Synthesis	AR
	L20	Dialogue Systems	AR

What you should know !

- **Mathematics**
  - Vectors and matrices
  - Calculus
- **Signal processing**
  - Speech signal processing (Speech Processing)
- **Probability theory**
  - Discrete and continuous probability
  - Joint probability and Bayes Theorem
  - Gaussian probability density functions
- **Programming**
  - Python
- **Linguistics and Phonetics**
  - Nothing (mostly)!

Does this coincide with your opinion ?

Some recommeded reading

- Daniel Jurafsky and James H. Martin (2018 draft). *Speech and Language Processing*, Pearson Education (3rd edition)
- Mark J.F. Gales and Steve J. Young (2007). *The Application of Hidden Markov Models in Speech Recognition*, Foundations and Trends in Signal Processing, 1 (3), 195-304
- Geoff Hinton, et al (2012). *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal Processing Magazine, 29(6):82-97
- Dong Yu and Li Deng (2014) Automatic Speech Recognition: A Deep Learning Approach.

This Lecture

- Introduction into Speech Technology

Why Speech Technology ?

Speech processing aims to model and manipulate the speech signal to be able to transmit (code) speech efficiently; to be able to produce natural speech (***synthesis***) and to be able to ***recognise*** the spoken word.

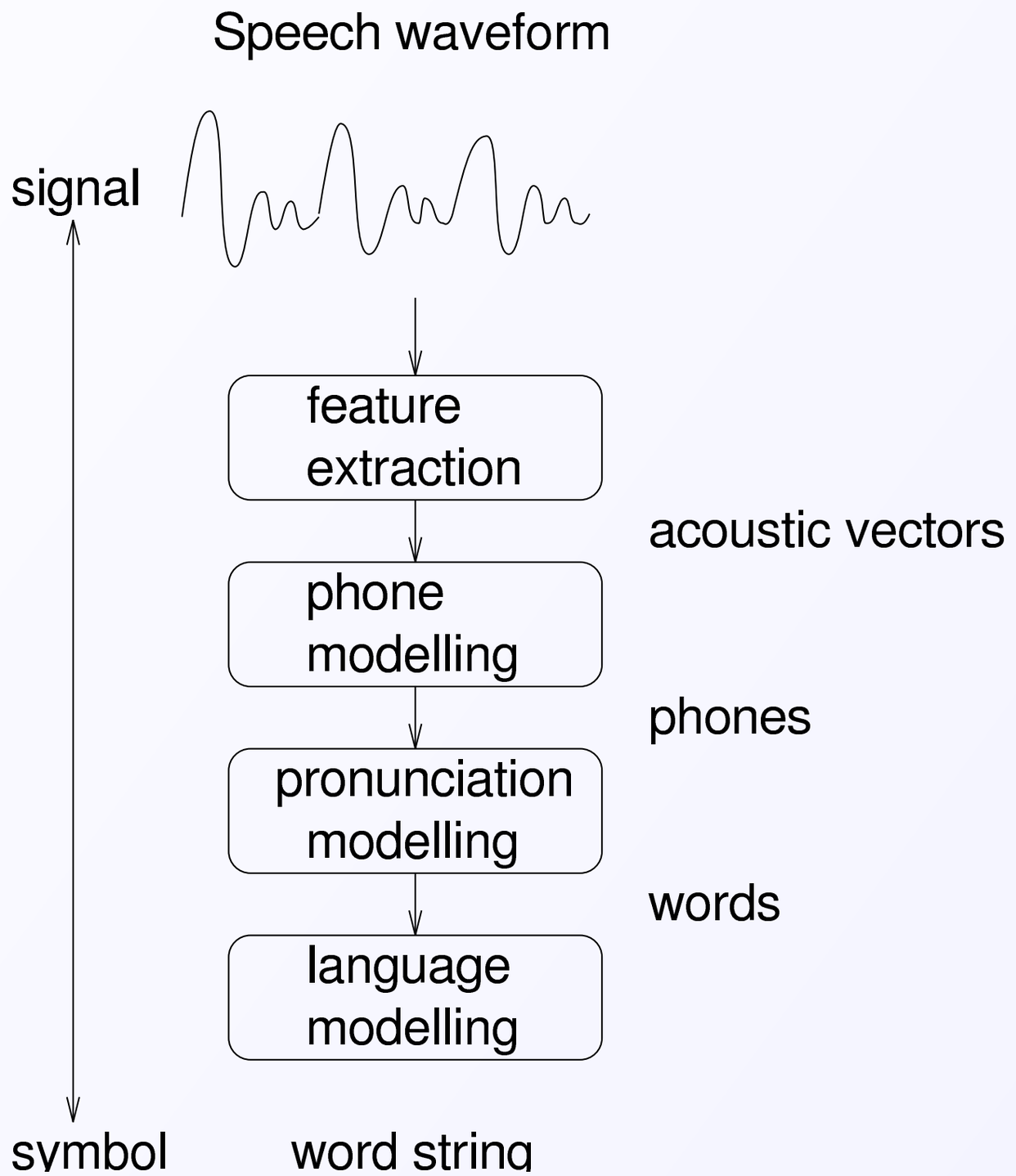
Since speech is the natural form of communication between humans it reflects a lot of the *variability and complexity* of humans! This makes modelling speech an interesting and difficult task.

The speech signal contains **information from many levels** and encodes information about the speaker and the acoustic channel; the words and their pronunciation; the language syntax and semantics etc.

Speech technology is becoming increasingly well

Levels of representation

Speech processing is mot often concerned with transformation of representations at different levels.



Stepping up the levels

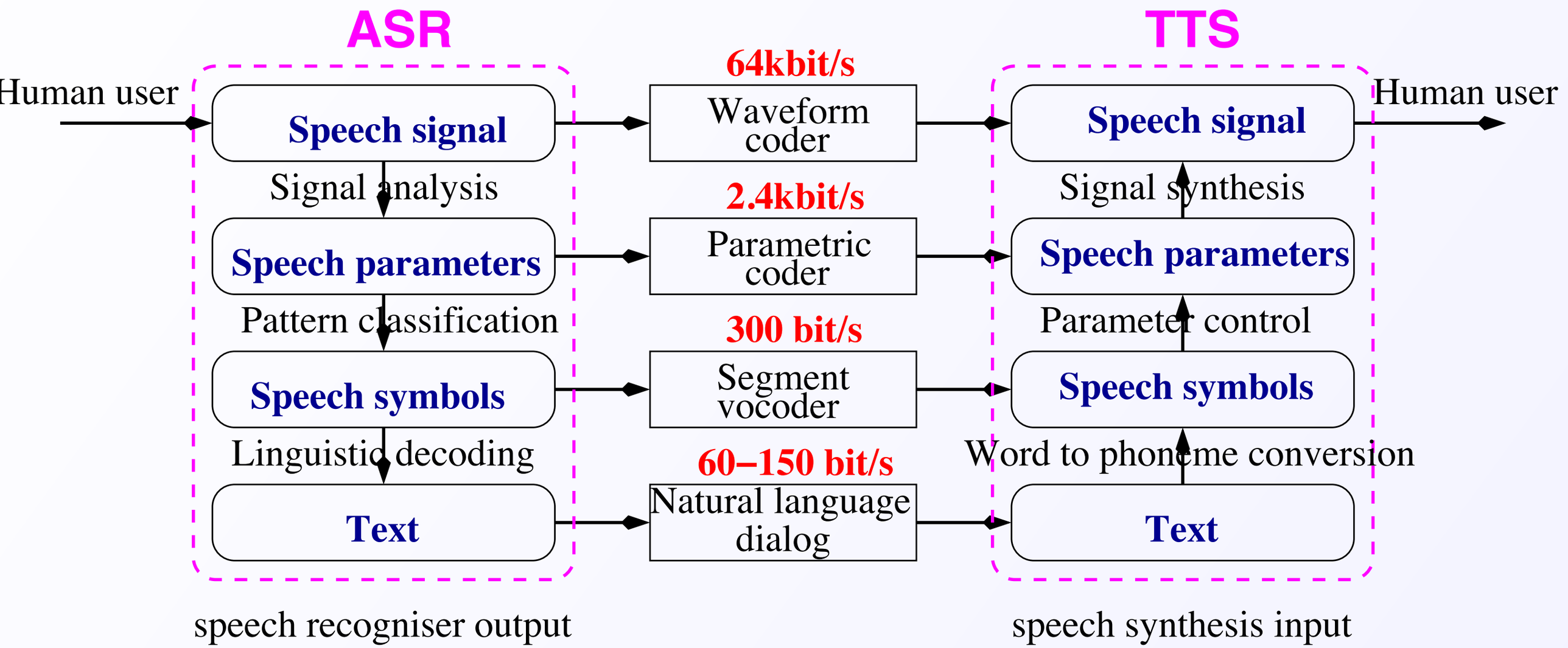
Speech technology requires to extract information from the speech signals at different levels. Here are some examples sarting from lower levels

1. **Low - no phonetic content**
   Signal to noise ratio estimation, pronunciation quality measurement, pitch extraction, noise suppression
2. **Medium - no semantic content**
   Phone recognition, pronunciation training, keyword spotting, "Who spoke when", emotion detection, speaker identification, language identification
3. **High - no discourse or pragmatic content**
   Speech recognition
4. **Natural language processing**
   Summarisation, dialogue systems

At low levels quantities are continuous, at higher levels they are discriete !

Transfer of inofrmation

Speech signals can be modelled at different levels of complexity. The representations encode content as well as speaker specific information. Text is a simple and low bit-rate form to encode a speech signal.



Speech technology

… is everywhere

- Phones
- Desktops
- Fighter Jets
- Cars
- Lifts
- Toys
- …

...is big business

Tencent 腾讯

Google

LG

Alibaba.com

JD.京东 COM

CISCO

facebook

NUANCE

CISCO

HUAWEI

Baidu 百度

at&t

TOSHIBA

Microsoft

amazon

ITSLanguage

SAMSUNG

IBM

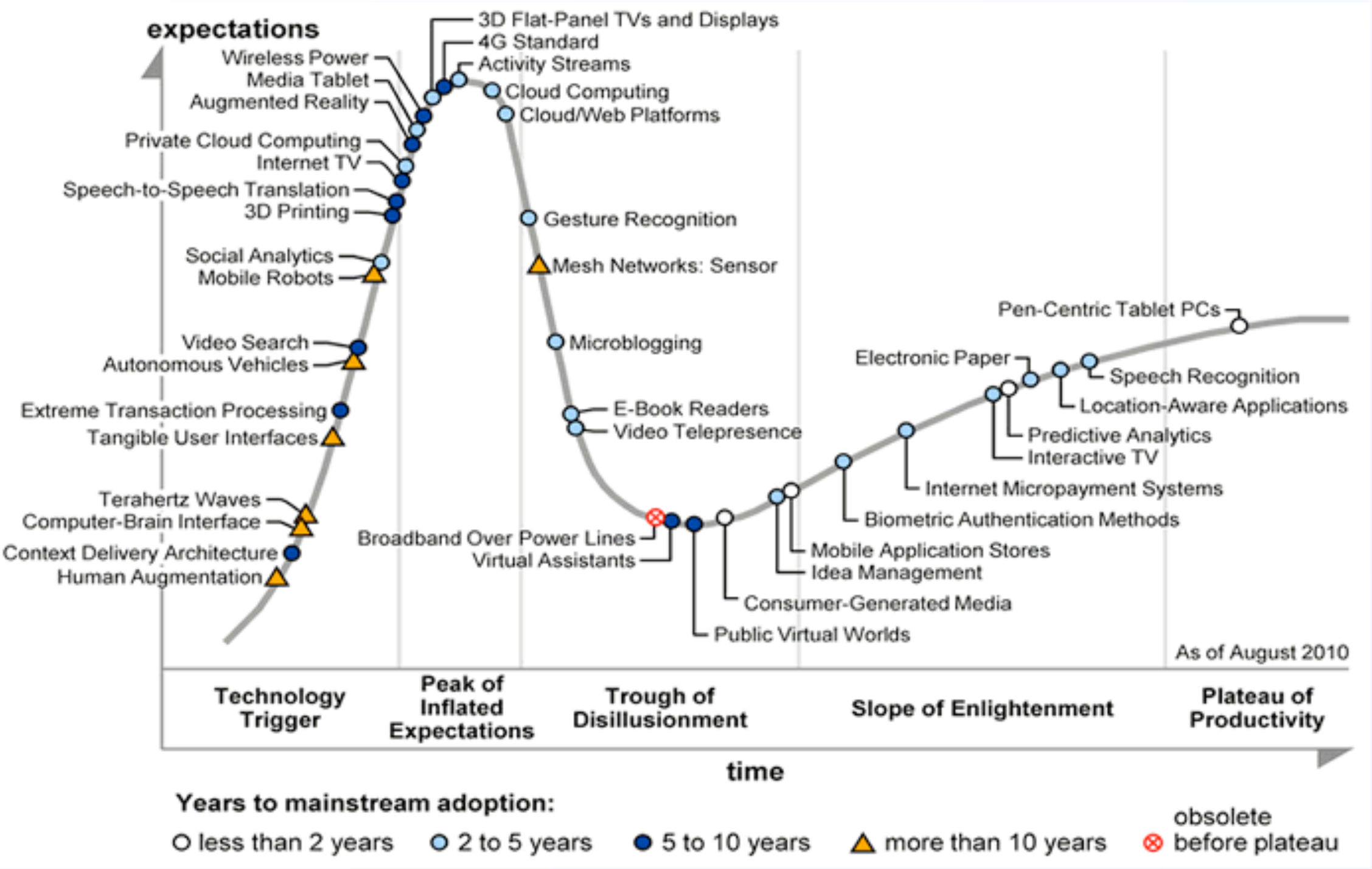
CIRRUS LOGIC

scribeTECH

Every high-tech IT company has now speech divisions

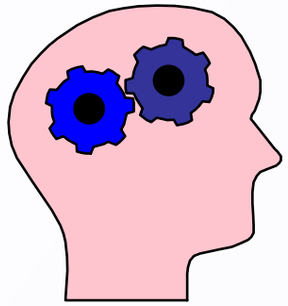
Often > 100 employees

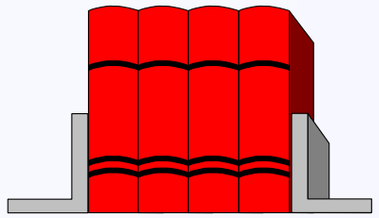
… is on the plateau of productivity



The Hype Cycle

.. is multidisciplinary





ARTIFICIAL INTELLIGENCE

Computational Linguistics

LINGUISTICS

Natural Language Proc.

Cognitive Science

Pattern Processing

SPEECH PROCESSING

Psycho-Linguistics

Information Retrieval

Dialogue

ENGINEERING

Human-Computer Interaction

PSYCHOLOGY

… is not "solved"

- Modern speech technology requires vast computing resources
- It is the expertise of companies to make that more manageable

It’s about talking to others

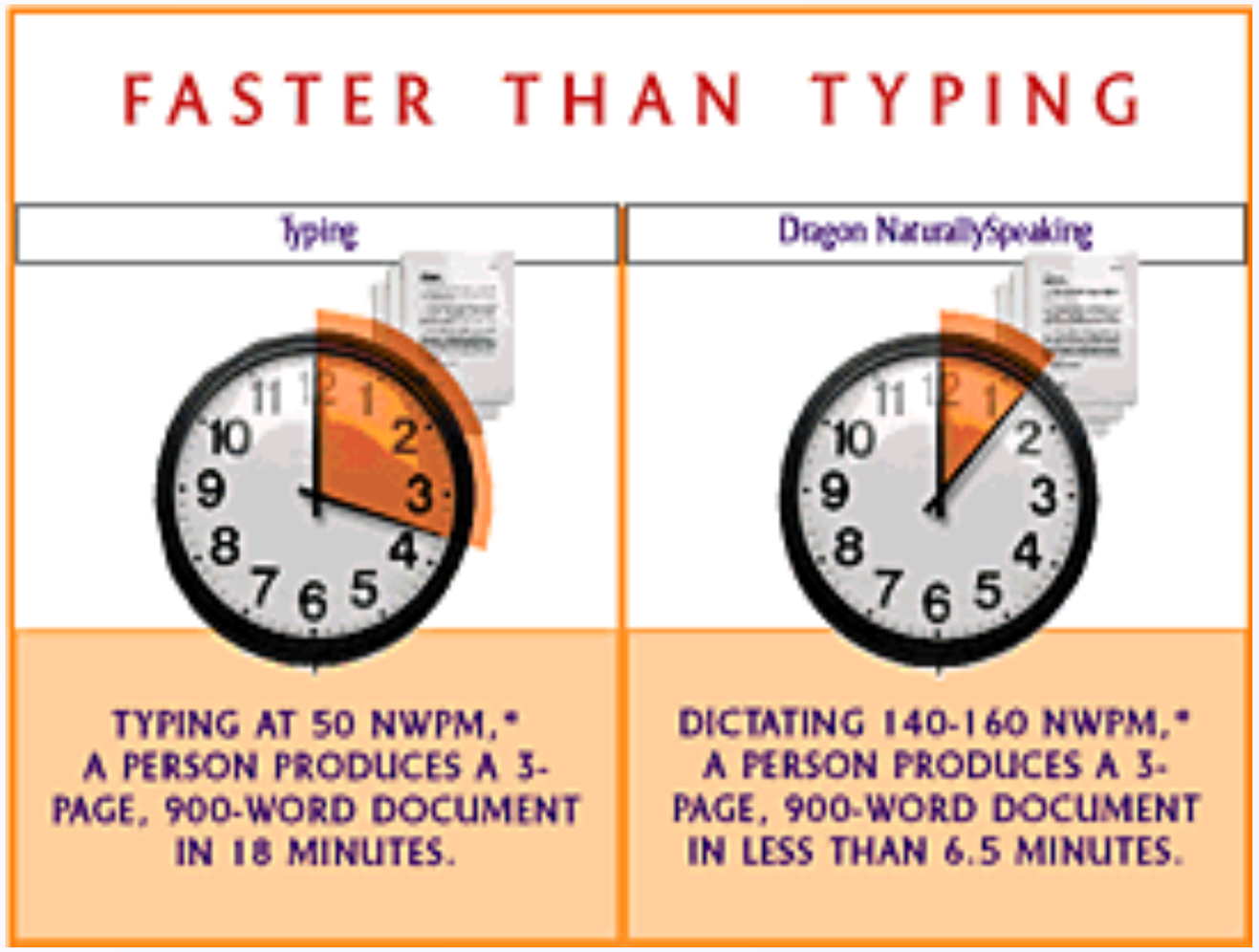
Human-Human Communication

1876: 2

2006: 1,000,000,000

It’s about talking to machines

Human-Machine Communication



It’s about listening to machines

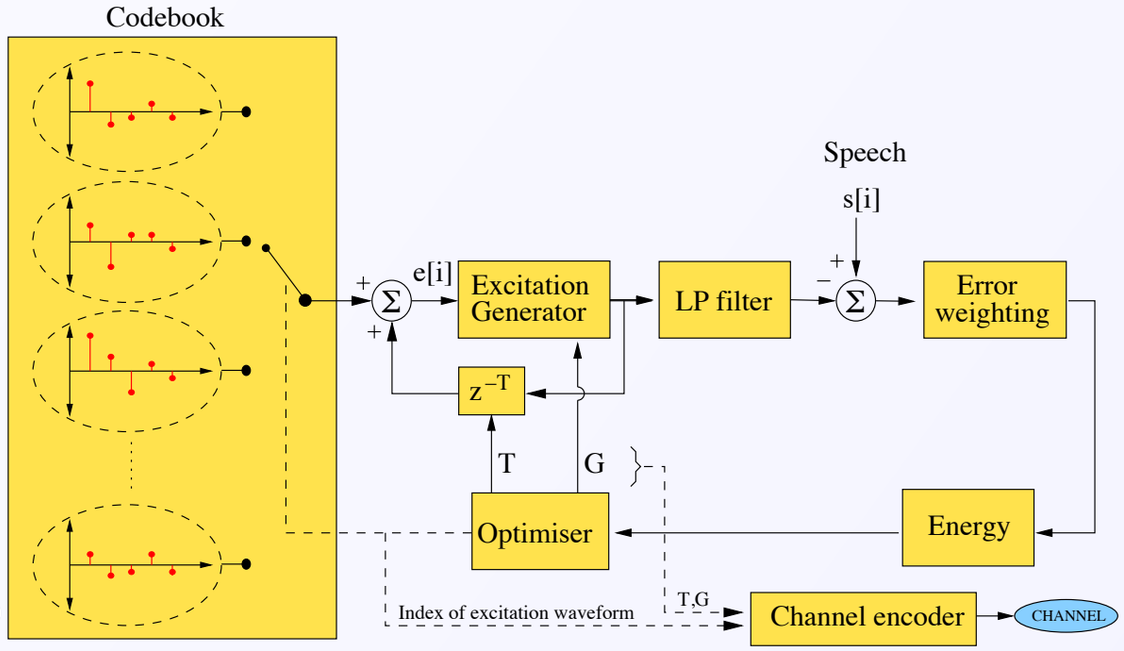
Machine-Human Communication

What are the technology tasks ?

- Core tasks
  - Compression
  - Enhancement and Transformation
  - Generation
  - Recognition
- Integrative tasks
  - Dialogue
  - Speech to Speech Translation
  - …

Compression

- Speech coding
  - essential for mobile phones
  - scaleable by bit-rates
  - latency
  - beween 300 bit/s - 32kbit/s
- Key paradigms
  - analysis by synthesis coding
  - AM coding
  - waveform coding

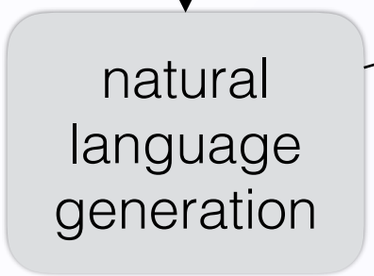


Enhancement and Transformation

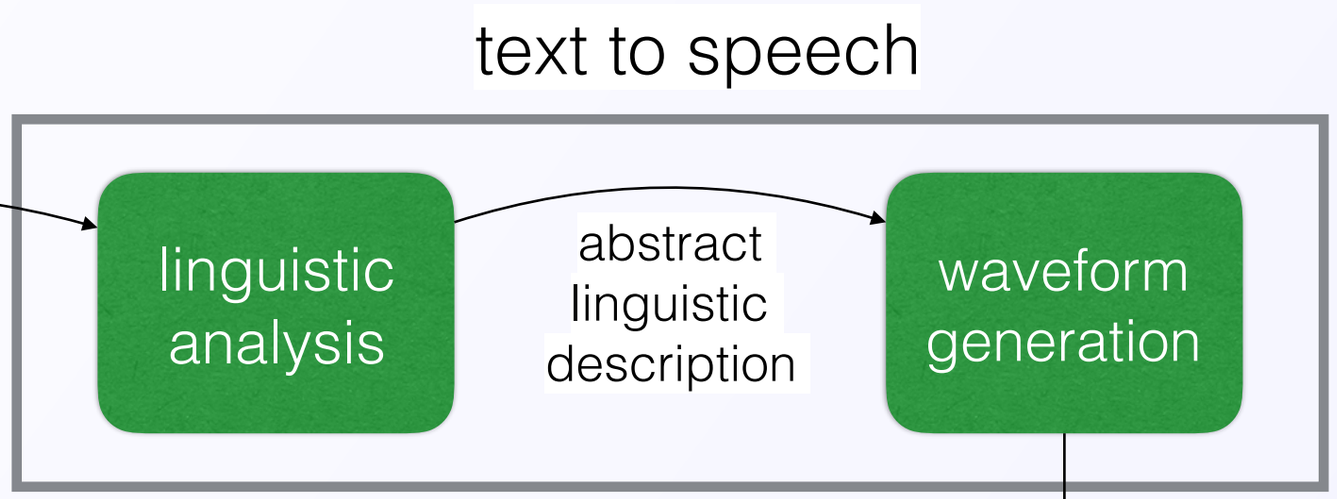
- Speech in noise
  - Essential for modern mobile phones
  - Source separation
  - Assessment using intelligibility and naturalness scoring methods
- Transformation to target speakers
  - Dubbing
- Key paradigms
  - Wiener filtering
  - Adaptive filtering
  - Independent component analysis (ICA)
  - Beamforming
  - Frame transformation

Speech generation and modification

concept



formatted
text

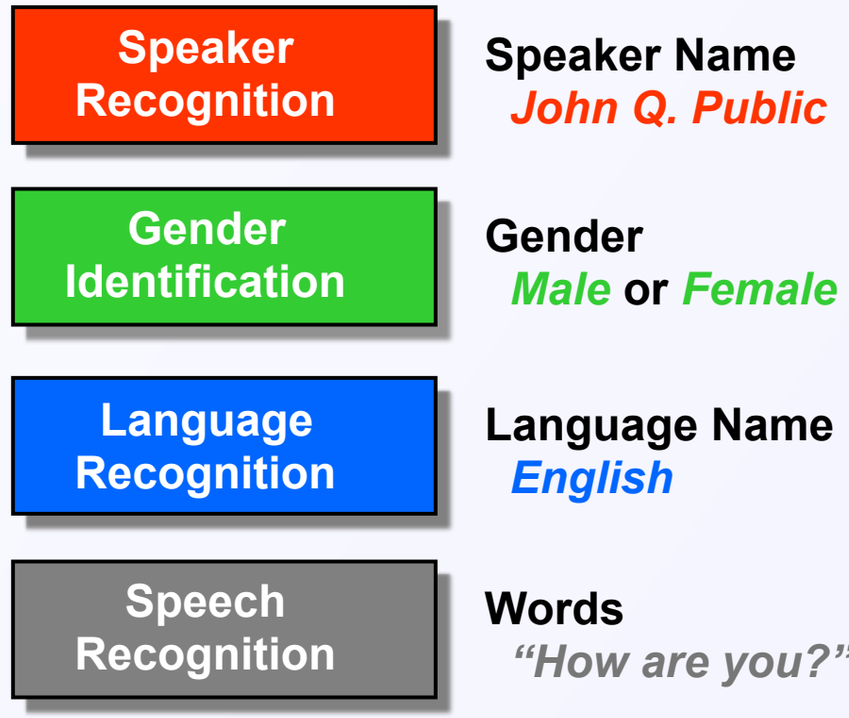


synthetic
speech
waveform

Recognition

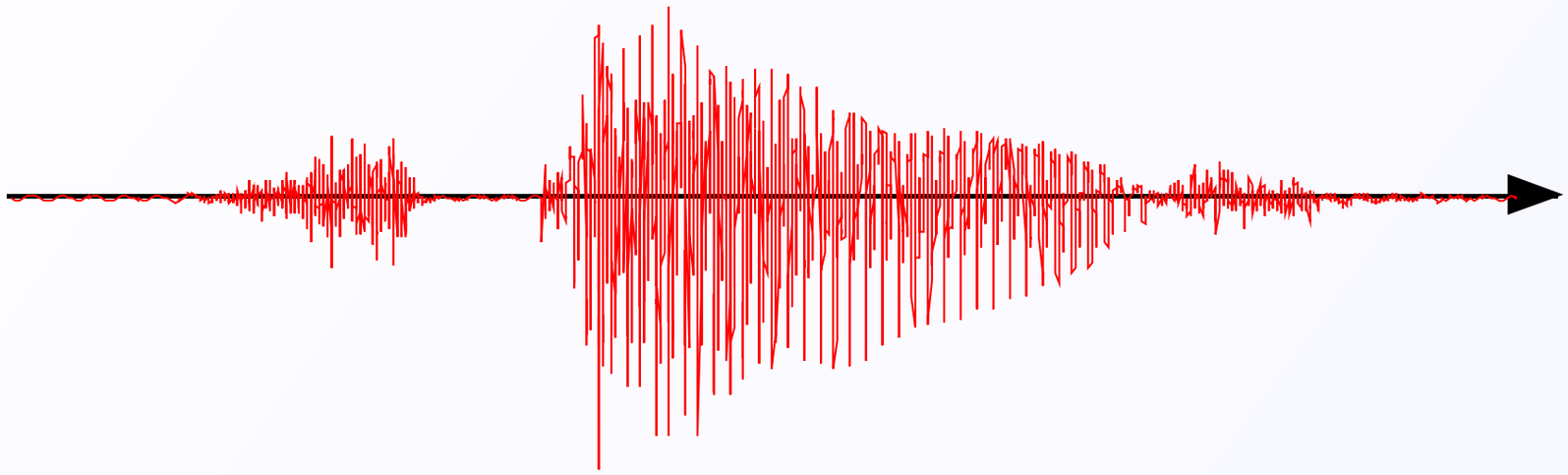
Classification is the task of assigning a sample to a particular class in a finite (!) set, for example a phoneme, a word, a speaker ! Hence most speech technology tasks use classification techniques:

- Automatic Speech Recognition (ASR)
- Speaker Identification and Verification
- Language and Accent Identification
- Emotion Detection
- Diarisation
- ...



The most complex of these tasks is ASR, and most other applications derive their algorithms from ASR principles.

What is speech recognition ?



The following distinction is usually made:

- **Recognition**
  Identification of the words in an utterance (speech to orthographic transcription)
- **Understanding**
  identification of utterance meaning.

How far can you go in building ASR (automatic speech recognition) systems without understanding? (A long way ...)
This course only deals with speech recognition.

General approach to speech recognition

Originally (1970s) a difference in philosophy to approach ASR:

- ***Recognition***
  Based on pattern matching techniques. Syntax/semantics learnt from statistics and not separated. Implicit knowledge learnt from (lots of) data. Stochastic systems use powerful algorithms to automatically optimise formal mathematical models for a given task.
- **Understanding**
  1970s style understanding systems were based on extraction of "perceptually important" features and the use of linguistic rules. Much use was made of explicitly coded syntax, semantics and pragmatics. Deterministic and rule-based. Many, interacting and ad hoc rules.

The traditional speech understanding approach is no longer used since it gave relatively poor performance and the capabilities of speech recognition grew. All modern speech understanding systems use text produced by a speech recognition system (sometimes multiple hypotheses) and then do further text interpretation. The current dominance of statistical models using parameters estimated from large corpora has also influenced computational linguistics where there is

Why is it difficult to recognise speech ?

Speech is a complex combination of information from different levels (discourse, semantics, syntax, phonological, phonetic, acoustic) that is used to convey a message.
The signal contains much **variability** (important difference or noise?).

- **Intra-speaker**

  physical/emotional state, environment , etc...

- **Inter-speaker**

  physiological, accent/dialect, etc...

- **Speaking style**

  read/spontaneous, formal/casual

- **Acoustic channel**

  record utterance and noise, telephone channel, background speech, noise, etc...

Variability

ASR devices often lump together many of the variability sources. An ASR system needs the means of dealing with (i.e. capability to model)

- **Spectral variability**

  Linear or non-linear effects due to all variability sources

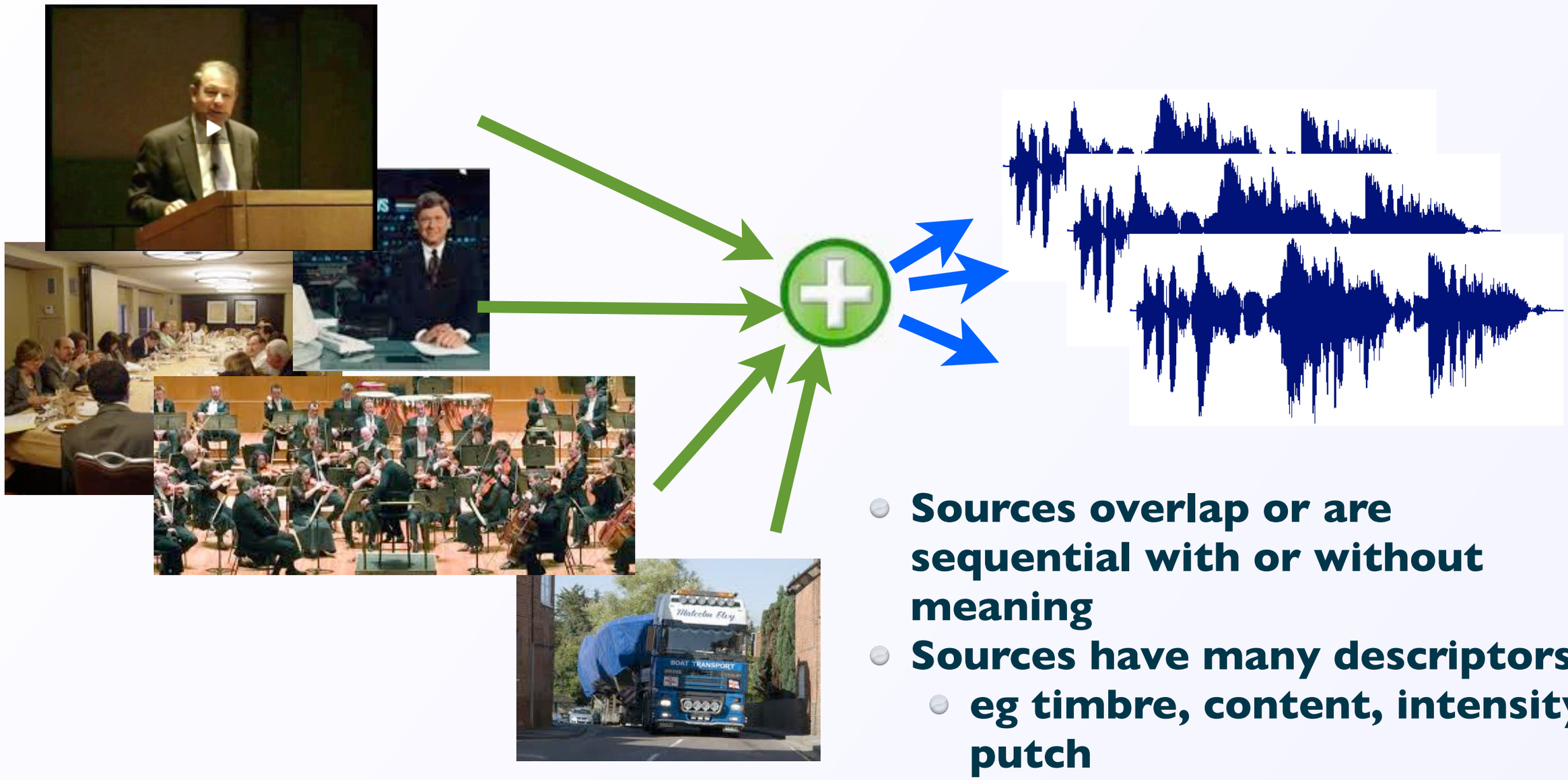
- **Timing variability**

  Mostly non-linear effects, speech can be stretched in non-linear fashion. More variations for speaker independent and continuous speech

The importance of effects varies with the tasks.

Speech in natural environments

Many sound sources present - eg Youtube



- **Sources overlap or are sequential with or without meaning**
- **Sources have many descriptors**
  - **eg timbre, content, intensity, putch**

Speech recognition - Task classification

Tasks are typically constrained to limit the variability by e.g.

- isolated word format
- limited vocabulary
- constrained syntax
- low noise condition
- etc...

Research tends to focus on making recognition systems **more general** ("all purpose"): Large vocabulary speaker independent continuous speech recognition systems trained on data from a variety of different sources.
Recogniser capabilities can be defined along a number of dimensions.

- **Speech**
- **Environment**
- **Linguistic criteria**
- **Input/Output**
- **Internal specifications**
- **Platforms**

Meetings

We spend a lot of time in meetings

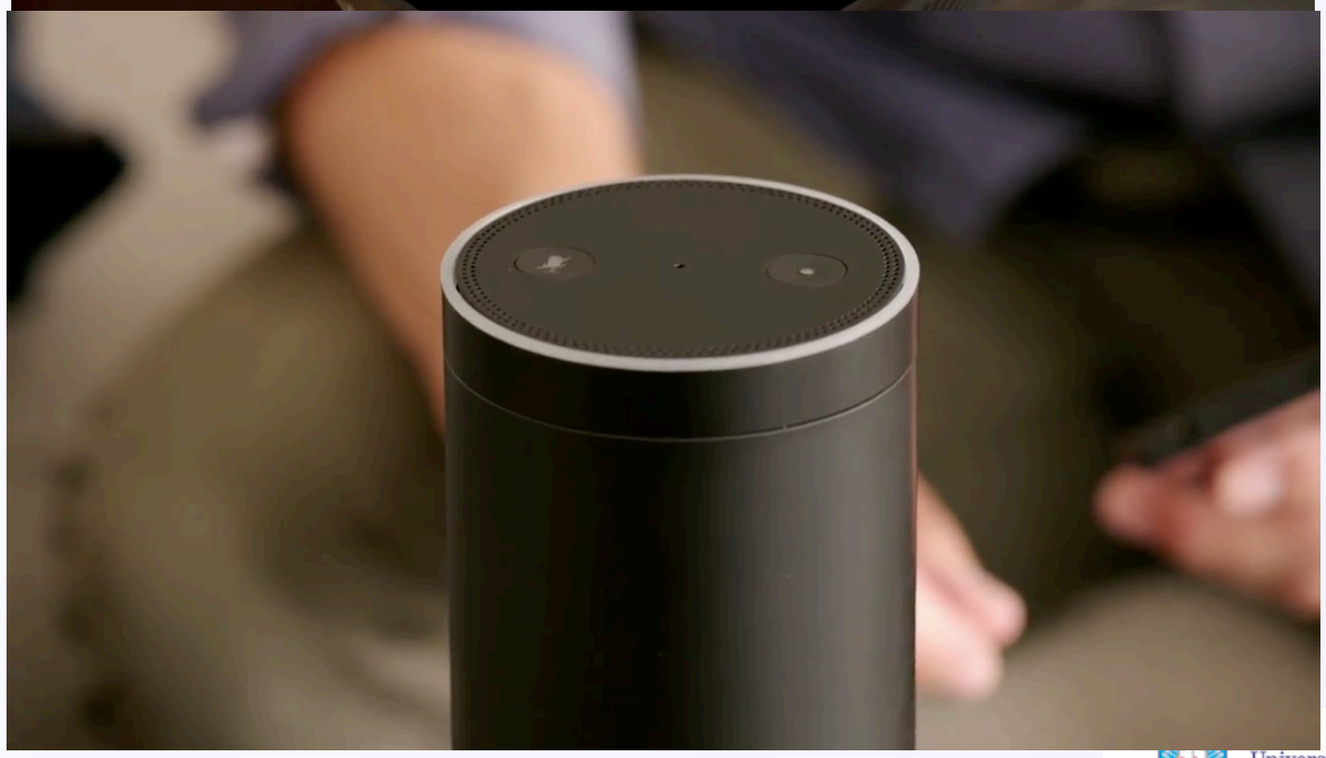
Language Learning

Speech technology in your living room

2011 KINECT™ for XBOX 360

2014 amazon echo

2015 jibo



University Of Sheffield.

System Aspects

What makes a "good" ASR system ?

- **Low error rate**

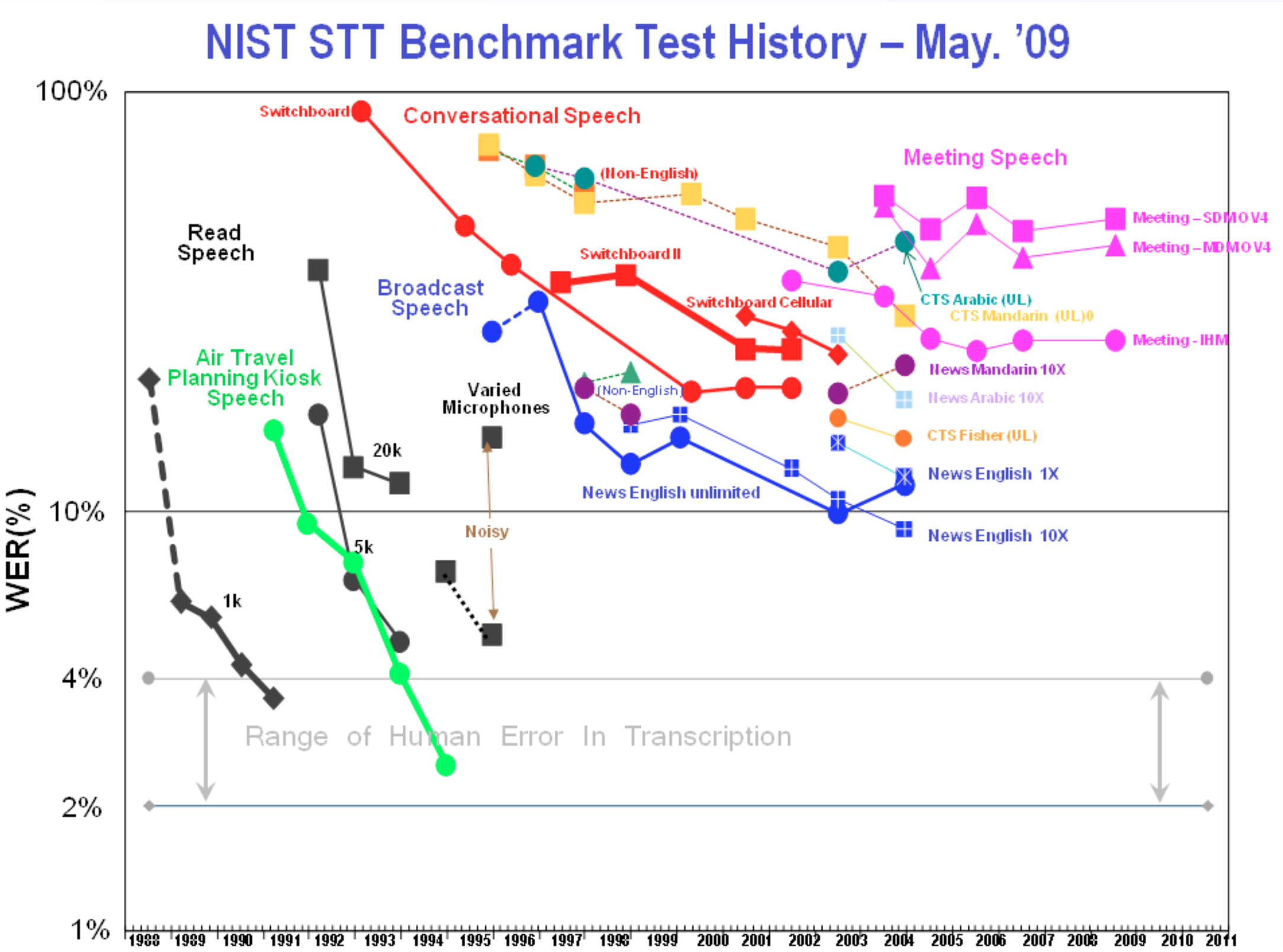
  Performance of speech recognisers can be measured by comparison of the output string with a manually transcribed version (reference transcript).

- **User satisfaction**

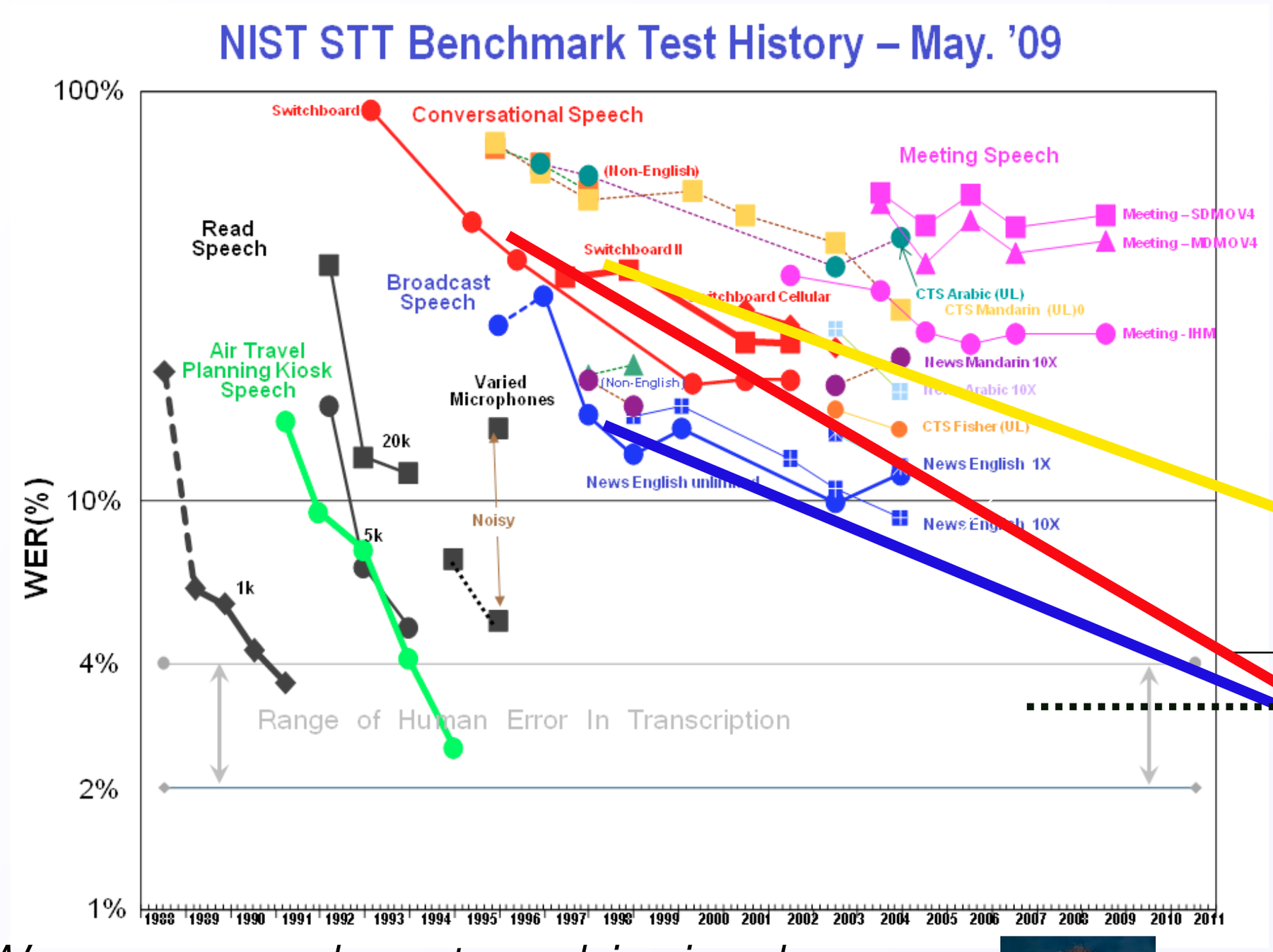
  Recognisers form part of a larger system (e.g. a text input system or an enquiry system). Users (customers!) are interested in overall system performance (e.g transaction time) rather than the raw error rate.

It is necessary to integrate the recogniser (& its shortcomings) into system design so the system can cope with recognition errors, e.g. by use of confirmatory strategies or the design of user interfaces to **allow correction** (i.e. the recogniser must be able to provide alternative recognition hypothesis strings, scores of confidence, addition of new words/phrases, ...)
ASR systems have a history of reporting worse performance in field with real users than in lab (different noise conditions, different user behaviour etc.). Hence it is very important to collect **realistic databases** for system development/test.

Progress and transferability



How do we do now ?



IBM 2017

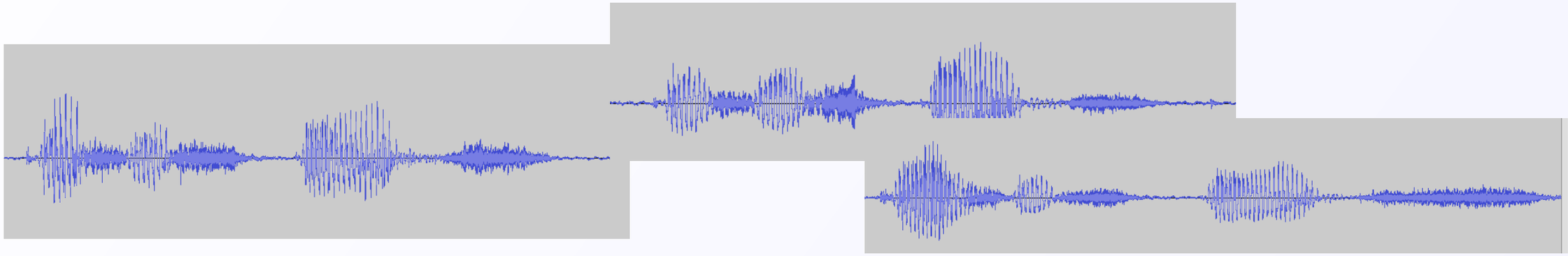
We are very close to achieving human transcription performance on conversational [..] speech.

David Nahamoo
Speech CTO at IBM Research

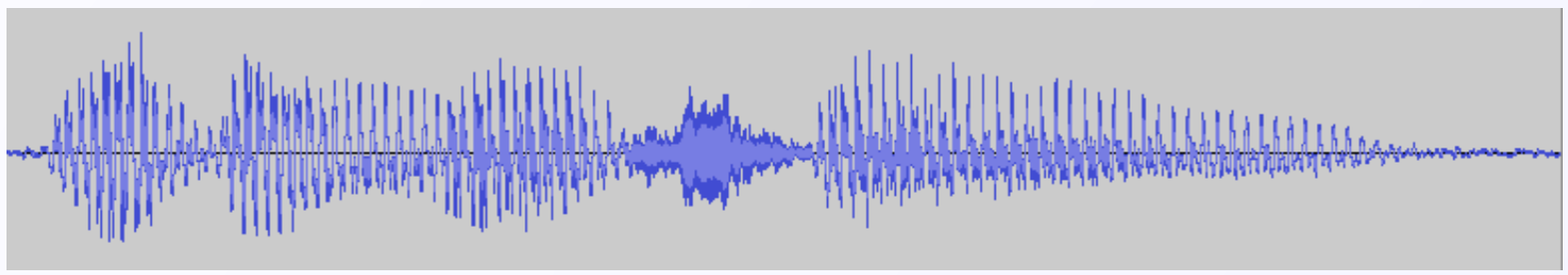
Classification of speech sounds

Speech recognition is about finding out what words where spoken

- We need to have a collection of examples



- Any unknown sound can be compared against the reference sound



?

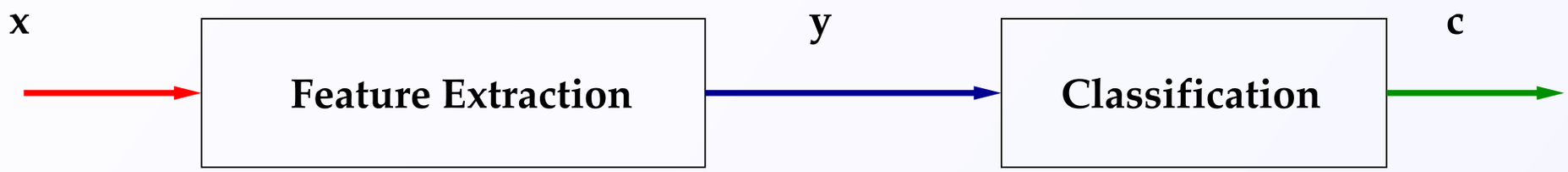
Classification of speech sounds (2)

- The first speech recognition systems worked on this simple framework. First recordings are of many words are obtained., together with a written form of what was said. These form the example set (often also called training set).
- During recognition, when a new audio file arrives, the recognition process involves comparing the new audio file with all other audio files in the example set. That comparison involved computation of a **score** (distance).

- You can produce an example set yourself. Good recording software is available at http://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

Pattern classification - General

Pattern classification is one of the most important tasks in pattern processing. The task is to identify patterns in the input data such that the data can be assigned to one out of a finite number of classes C_i.
All pattern processing techniques make use of the fundamental **pattern classification paradigm.**



Example **Representative Vector** **Class**

The process of classification of patterns always operates in two stages

1. Feature extraction
2. Classification

A single example x is presented to the system, essential information in the form of a vector y is extracted which is then classified into class C.

Statistical speech recognition

In ASR the task is to find the most likely sentence (word sequence) W used for generating the utterance A.

$$\hat{W} = \arg\max_{W} P(A|W)P(W)$$

The essential parts of an ASR system are

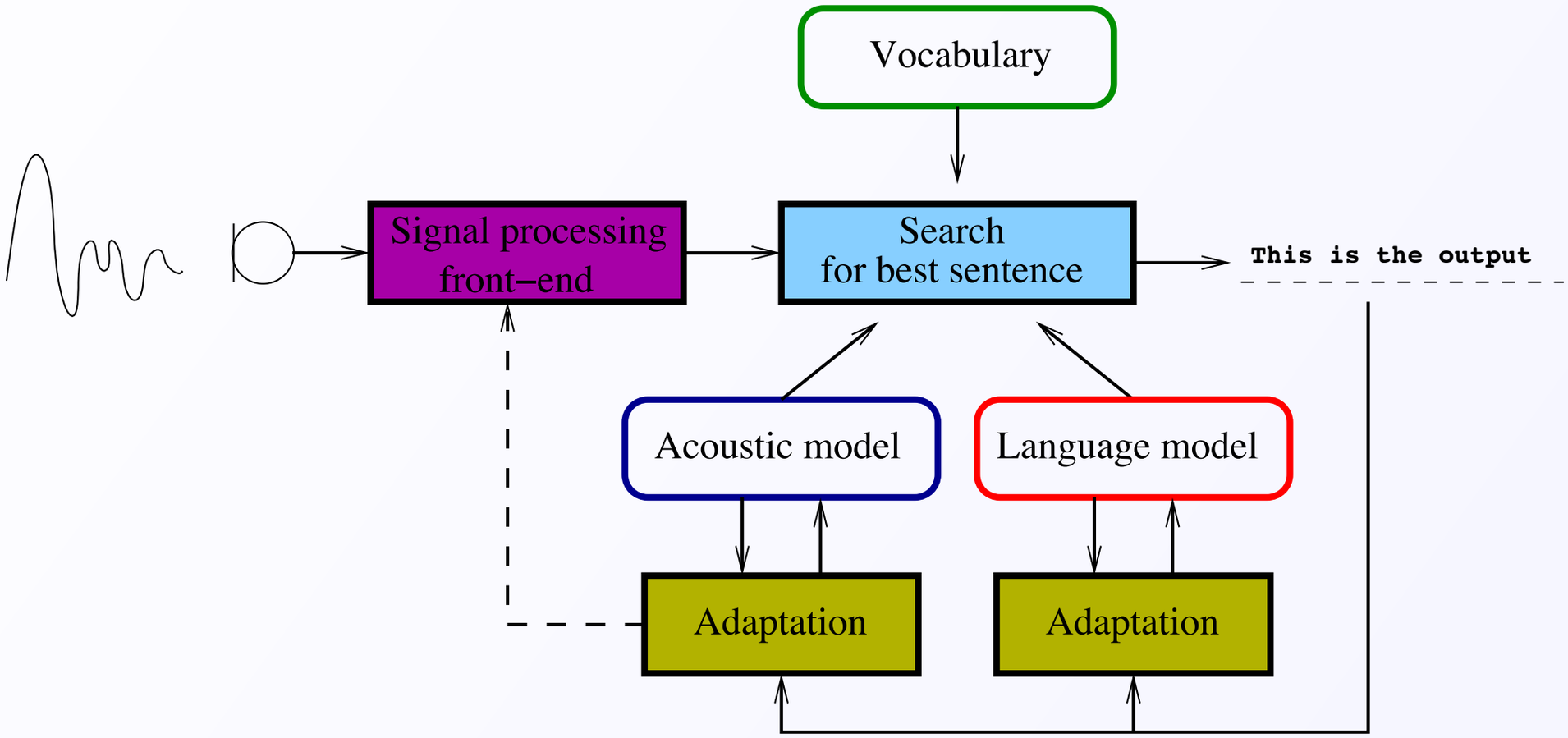
- ➜ the **acoustic model** $P(A|W)$
- ➜ the **language model** $P(W)$

The ASR problem depends on finding solutions to two problems

- ➜ **Training**
  finding suitable representations for the acoustic model and the language model.
- ➜ **Recognition**
  finding the most likely word sequence $\hat{W}$

The most common form for the acoustic model is a **hidden Markov Model** (**HMM**). The most common form for the language models are N-grams.

Generic Recognition Architecture



A search is made for the most likely word or sentence given the acoustic and language models (recognition,decoding). A finite set of words is defined in the vocabulary of the ASR system.

- ➜ The **front-end** converts the audio stream into a stream of feature (observation) vectors.
- ➜ The **acoustic model** is responsible for matching acoustics and individual words as defined in the vocabulary. This is the most complex part of an ASR system.
- ➜ The **language model** represent syntactical, semantical, discourse constraints. In some cases no (a null) language model is appropriate.
- ➜ The output of the ASR system may be used to adjust the models (acoustic or language model **adaptation**).
- ➜ Words not in the vocabulary (Out-Of-Vocabulary) cause errors.

Isolated word recognition simplifies the system by pre-segmentation of the speech data. However, this is not straightforward in anything but a low-noise audio environment.