# Comparing human and automatic speech recognition

Odette Scharenborg

*Centre for Language Studies*
*Radboud University Nijmegen*
*The Netherlands*

o.scharenborg@let.ru.nl
http://www.odettes.dds.nl

***Based on:*** *Scharenborg (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. Speech Communication.*

## Comparing…

… aims, focus, and research approaches

… ‘implementation’ of the (word) recognition processes

… speech recognition performance

Concluding remarks

# The speech recognition process

Investigated by
- Human speech recognition (HSR)
- Automatic speech recognition (ASR)

**Central issue:** Word recognition

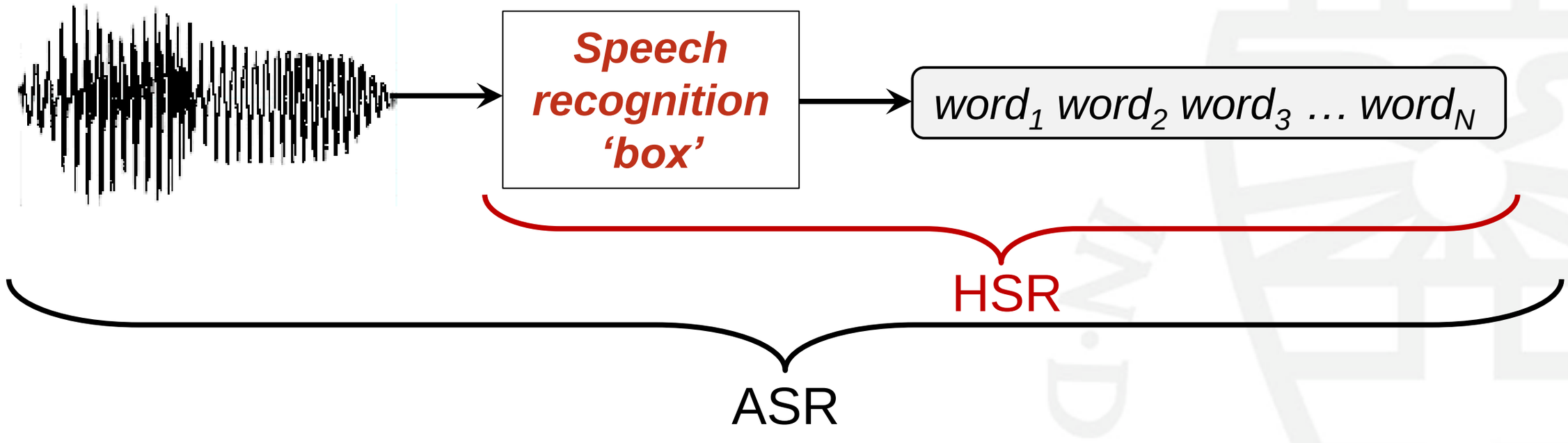
However, **different**
- Research aims
- Research focus
- Research approaches

# Different aims

**HSR:** Understand how listeners recognise spoken words

**ASR:** Build algorithms that are
able to recognise words automatically,
under a variety of conditions,
with the least possible number of recognition errors

# Different focus




ASR: Algorithms completely understood mathematically

HSR: No unified theories accounting for all aspects of human spoken word recognition; many parts of the theories remain unspecified

# Different research approaches

**HSR:**

- Behavioural studies
  - (Auditory) lexical decision
  - Gating
  - Phonetic categorisation
  - Eye-tracking
- Brain studies, (f)MRI, EEG
- Computational models

→

Reaction times
Phoneme response probabilities
Error/identification rates
Eye movements
Brain waves
Predictions

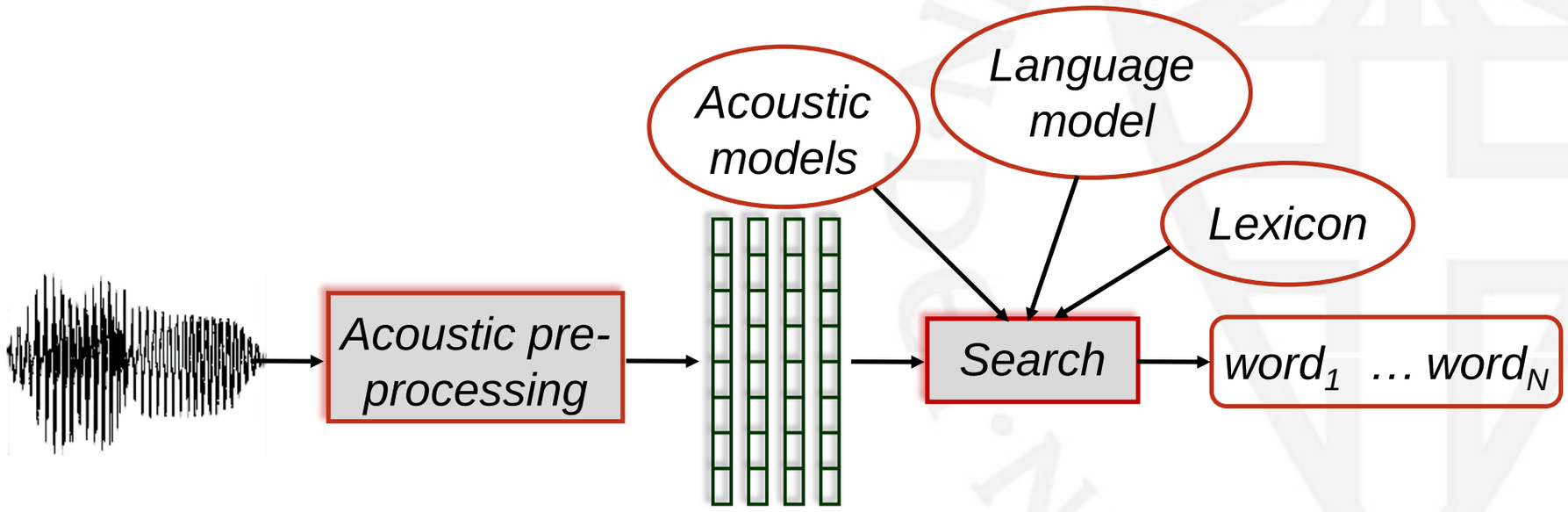
↓

Theories on (parts of the) human speech recognition process

# Different research approaches

**ASR:** Improvement of, e.g.:

- Signal representations
- Search methods
- Robustness in adverse conditions
- Language modelling



# Comparing (word) recognition processes



Based on: Scharenborg et al. (2005). Cognitive Science.

# The invariance problem

= Map a highly variable acoustic signal onto discrete representations (e.g., words)

Two ‘extreme’ solutions:

1. Episodic theories of lexical organisation
    - Multiple stored acoustic representations for each lexical unit: HSR & ASR

2. Abstract representations:
    - HSR: prelexical and lexical levels
    - ASR: subword acoustic models

# Real-time processing

HSR:
- Necessary for efficient communication
- Continuous flow of information between the prelexical and lexical levels

ASR:
- Not always important in ASR applications
- Graded, continuous matching between the signal and the acoustic models

# Multiple activation and evaluation of words

1. Compare input with word representations in lexicon
   - Search through lexicon
   - Activate all (partly) matching words

2. Assess degree to which input matches word representations, in parallel

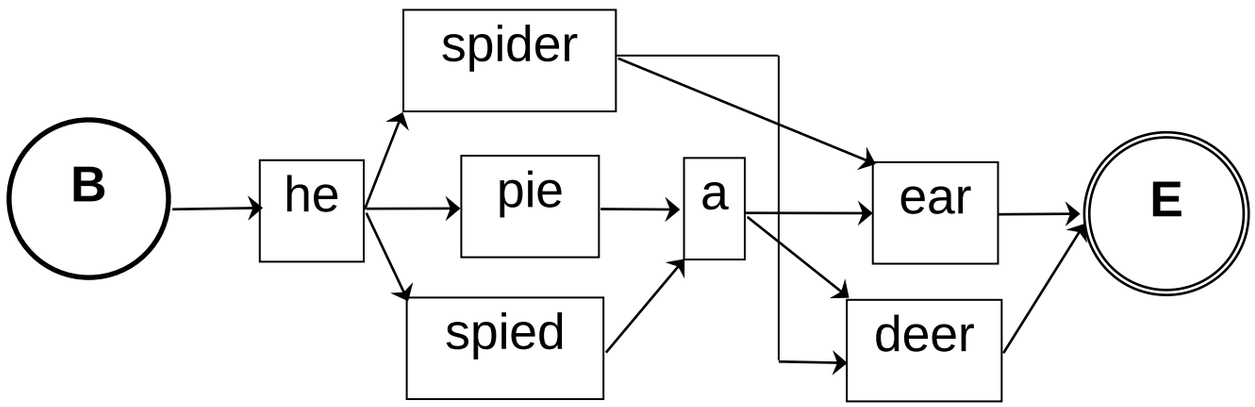
3. Choose best-matching word (HSR) or path (ASR)

HSR: subcomponents sometimes correspond to separable parts

ASR: subcomponents are generally combined in a single search process

# Recognition of continuous speech

ASR




HSR

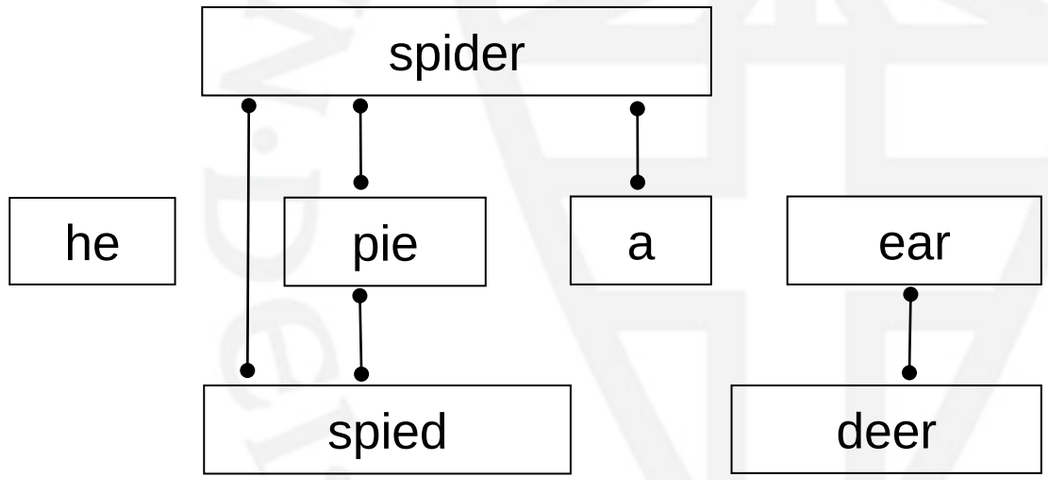





Figure 2 in Scharenborg et al. (2005). Cognitive Science.

# Cues to lexical segmentation

- Humans use word boundary cues, e.g.
  - Rhythmic structure
  - Phonotactic constraints
  - Acoustic and allophonic cues
  - Silent pauses

- ASR systems generally do not use these type of cues

# Comparing human and machine speech recognition performance

# The need for comparative studies

Investigate…

- the size of the ‘performance gap’
- why human speech recognition is superior to ASR
- what can be learned from HSR to improve ASR performance

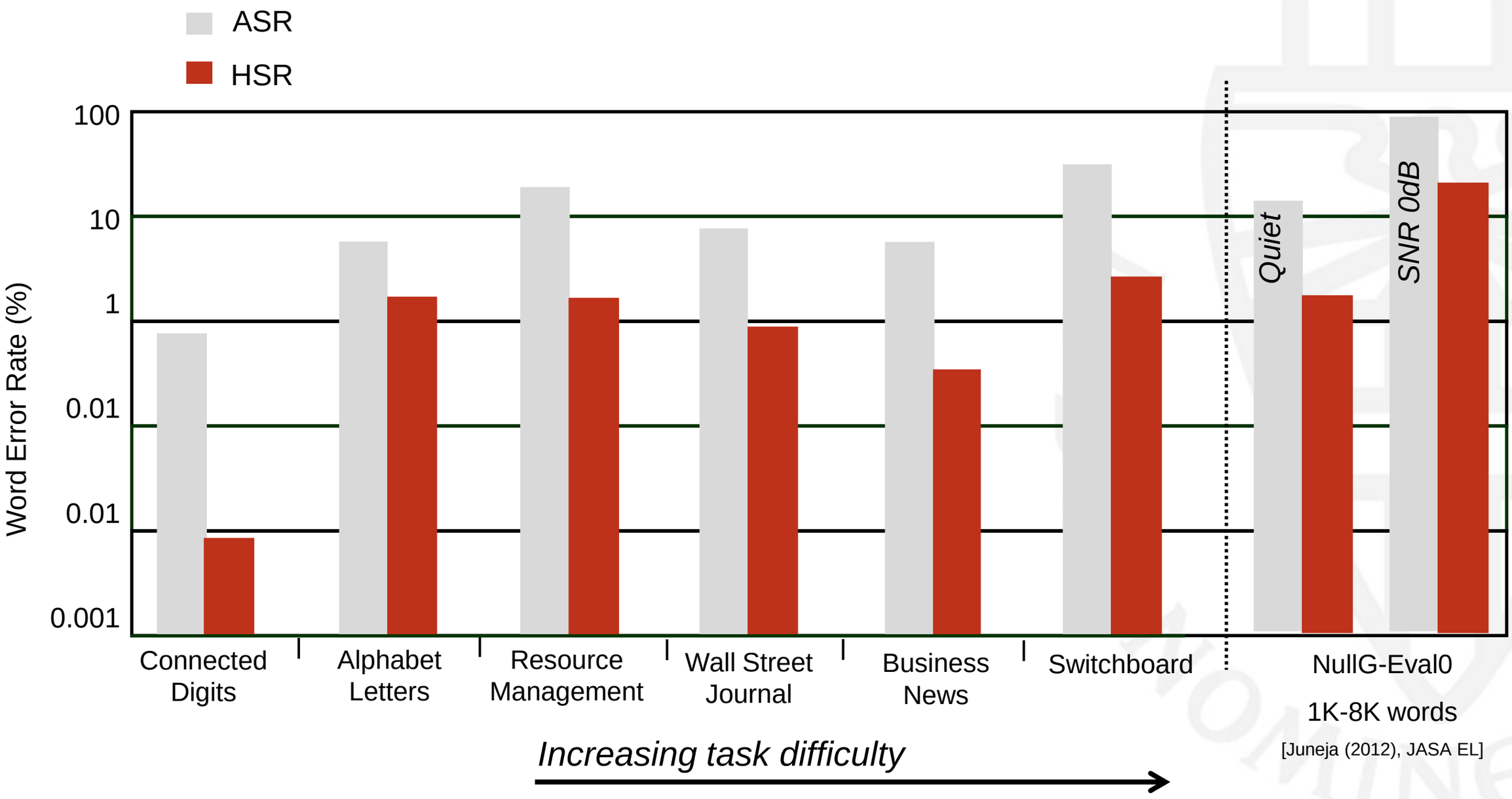
# Few comparative studies exist

- Task differences

- Measurement differences:
  - ASR: (word) error rates
  - HSR: reaction times / eye movements / (in)correct responses

- Differences in data sets:
  - ASR: large(r), dedicated data sets
  - HSR: small sets of dedicated stimuli

# Human vs. machine word recognition performance




**Source:** Lippman (1997). Speech Communication.

Figure adapted from: Moore (2003). Eurospeech.

# Findings comparing humans and machines

- Similar type of errors = similar type of recognition difficulties
- ASR systems just make more
- Content words more difficult than function words
- But humans make fewer inflection errors

- Performance gap increases in *adverse* conditions

# What causes this performance gap?

- Training material:
  - ASR: hundreds of hours
  - Humans: hundreds of *thousands* of hours [Moore (2001, 2003). Eurospeech]

⇨ Simply adding training material does not help (enough)

- Training material:
  - ASR: hundreds of hours
  - Humans: hundred *thousands* of hours [Moore (2001, 2003). Eurospeech]

⇨ Simply adding training material does not help (enough)

- Higher-level information:
  - Humans: knowledge about world, environment, topic of discourse, etc.
  - ASR: language models containing word (co-occurrence) statistics

# Human vs. machine phoneme recognition performance

- Oldenburg LOgatome (OLLO) speech corpus [Wesker et al. (2005). Interspeech]
  - Consonant recognition: humans >>> ASR systems with about 40% [Meyer et al. (2011). JASA]
- Interspeech 2008 Consonant Challenge Corpus [Cooke & Scharenborg (2008). Interspeech]
  - Consonant recognition: humans >>> ASR systems, especially for plosives
- Not all phonemes are equally difficult to recognise for machines

⇨ Even when higher-level information is removed, humans outperform machines

- Training material:
  - ASR: hundreds of hours
  - Humans: hundred *thousands* of hours [Moore (2001, 2003). Eurospeech]

⇨ Simply adding training material does not help (enough)

- Higher-level information:
  - ASR: language models based on statistics
  - Humans: more than just word frequency and word co-occurrence probabilities, priming

- Acoustic cues/features

# Humans' vs. machines' use of acoustic cues/features

- Entire speech signal vs. acoustic feature representation (MFCC / PLP / …)
- Information in the speech signal not extracted: Phase-information
  - Important for intelligibility [Alsteris & Paliwal (2006) .Speech Communication]
- Lower-level differences in cue use:
  - Voicing information: ASR < humans
  - Place of articulation: ASR = humans
  - Plosives and non-sibilant fricatives: ASR > humans

# How can ASR be improved?

- Add contextual information, e.g., through priming

- Improve feature extraction
  ⇨ Focus on those cues whose recognition performance can be improved

… through human-machine feature/phoneme recognition comparisons

- Improve robustness against adverse listening conditions
  - Note: relatively new field in psycholinguistic research

… through human-machine feature/phoneme recognition comparisons

# Concluding remarks / Summary

## HSR vs. ASR:

- Not many collaborations between the two fields

## Recognition processes:

- Many similaraties in the computations that need to be carried out in order to recognise speech

## Performance comparisons:

- Show that humans are more *flexible* than ASR systems
- Show where potential ASR improvement is still to be gained
- Helps identify which parts of HSR knowledge can be used to improve ASR systems