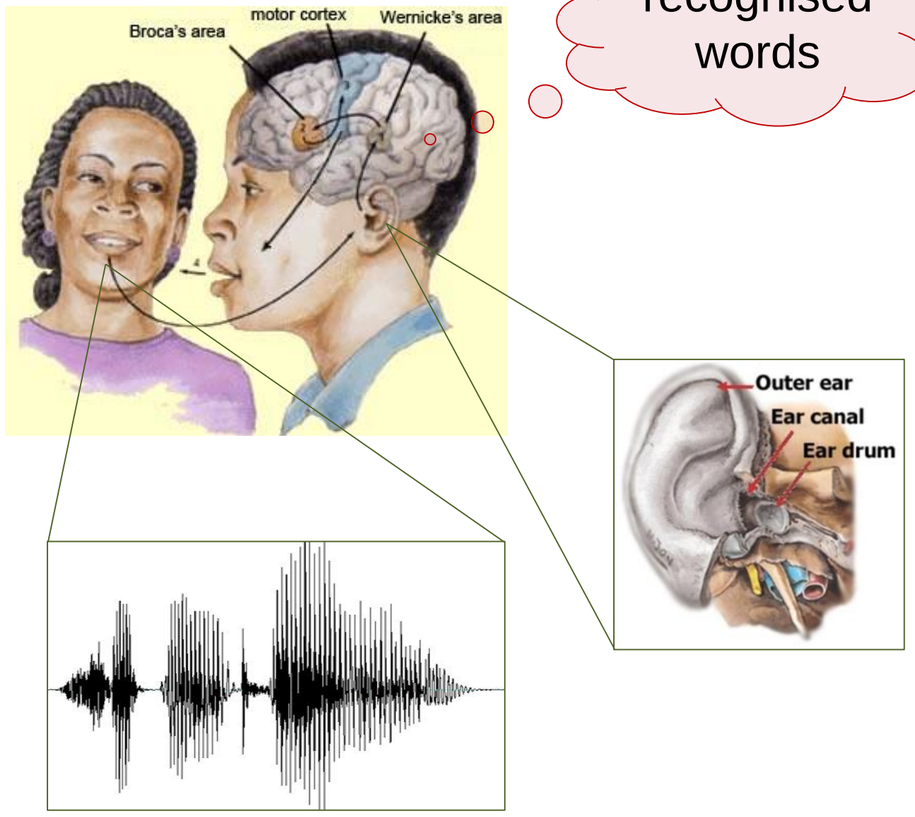
How do we recognise speech?

Speech perception in adverse listening conditions: Day 2

Odette Scharenborg
Centre for Language Studies
Radboud University
Nijmegen

E-mail: O.Scharenborg@let.ru.nl

Radboud University



Task for the listener:

Map the highly variable, continuous speech signal onto discrete units such as words

Overview today: Five key components of speech recognition

1. Multiple activation and evaluation of word candidates
2. Recognition of continuous speech
3. Theories of spoken-word recognition
4. Flow of information
5. (Flexibility of the perceptual system: Friday)
6. Models of spoken-word recognition
7. Depending on time: speech perception in a non-native language

Multiple activation and evaluation of word candidates

Radboud University

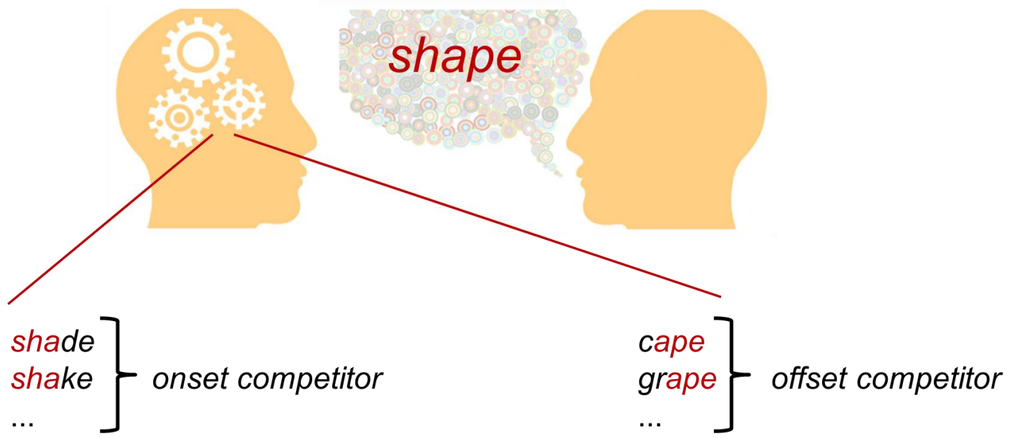
[Partially based on McQueen, 2005]

Multiple activation and evaluation of word candidates

Limited number of sounds in a language

A lot of phonological overlap between words

Many words resemble one another



/θə sʌn rɑɪzɪz/
/θə sʌnrɑɪzɪz/
/θə sʌnrɑɪz …/
/θə sʌn rɑɪ …/
/ … ɑɪ …/

All words that resemble the input are considered and evaluated in parallel

Word activation

- Activation level of lexical items dependent on match/mismatch with input:
  Larger match >> smaller match

⇨ **Graded activation**

- Mismatch too large? → Lexical item no longer considered to be a candidate for recognition

– Very fast

- Word-initial overlap >> word-final overlap [Allopenna et al., 1998]

- Phoneme overlap: 

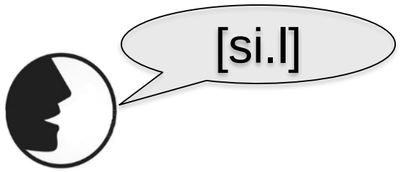
papal >> *pacer* [Connine et al., 1993, 1997]

Fine-phonetic detail modulates lexical activation

- Formant transition mismatches: *jog* >> *jo$_b$g* [e.g., McQueen et al., 1999]

- Durational information: [e.g., Davis et al., 2002; Salverda et al., 2003]
  At the syllable level: *ham$_{POLY}$ster* >> *ham$_{MONO}$ster*

  At the phoneme level: shorter [ʁ] → *dernier oignon* >> *dernier rognon* [Spinelli et al., 2003]

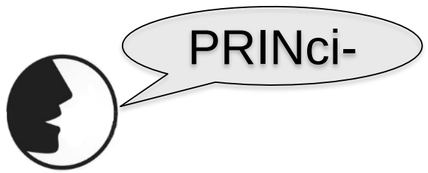
- Syllabic structure: 

si.lencio >> *sil.vestre* [Tabossi et al., 2004]

Other factors modulating lexical activation

- Phonological context: [grim] ≠ *green* [e.g., Gaskel & Marslen-Wilson, 1996, 1998, 2001; Gow, 2001]
  [grim] *paper* = *green*

→ Assimilation processes

- Word stress: 

PRINcipe >> prinCIpio [e.g., Soto-Faraco et al., 2001]

- Word frequency: high >> low [Dahan et al., 2001]

Evaluation of lexical candidates

Which word is recognised?

- With the best bottom-up goodness of fit?
- Competition and inhibition?

[but see Scharenborg et al., 2005; Norris & McQueen, 2008]

Speed and accuracy of word recognition

… depends on the number of activated words and their frequency of occurrence

Words are recognised **slower** and **worse** when they have

- Lots of neighbours:

  Bet → bed, vet, bit, bot, wet, let, bat, but, bell, beg

  Chaos → 0?

- Lots of neighbours with a high frequency of occurrence

⇨ Neighbourhood density? [Luce & Pisoni, 1998]

⇨ Size of the *word-initial cohort*?

[e.g., Hintz & Scharenborg, in prep; Magnuson et al., 2007]

Other factors that influence the speed and accuracy of speech recognition

Speech style (read speech vs. conversational speech; see yesterday)

Mother tongue of the speaker (native vs. non-native speaker)

Mother tongue of the listener (native vs. non-native listener; see later today/Thursday)

Background noise (quiet room vs. train station vs. pub; see Thursday)

Presence of a second task

Recognition of continuous speech

Radboud University

[Partially based on McQueen, 2005]

Recognition of continuous speech

= select the best-matching sequence of words for the acoustic input

While accounting for all speech - no *leftover* sounds allowed [Norris et al., 1997]

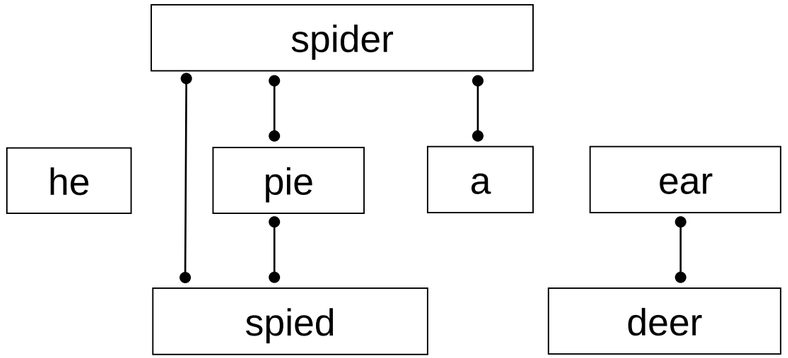


Figure 2 in Scharenborg et al. (2005). *Cognitive Science*.

Segmentation of speech into words

- Follows automatically from the best matching sequence

⇨ No explicit segmentation of the speech signal!

- Remember: no pauses between words (see Monday)

But….

Various cues signalling word boundaries exist

Word boundary cues

Phonotactic cues

- Phonotactics: [m#r] [e.g. Dumay et al., 2002]

- Probability of sound sequences [van der Lugt, 2001]

Metrical cues

- Rhythmic structure of speech, e.g., word stress on 1st syllable in English [Lots of work by Cutler]

Allophonic cues

- Aspiration of word-initial stops in English [e.g., Church, 1987]

- Sound duration (see earlier today)

Possible-word constraint (PWC)

- Word candidates that are misaligned with possible word boundaries are penalised
  → lower activation levels

- Misalignment: when a stretch of speech between the word and a likely word boundary does not contain a vowel

  f*apple* << ***vuff****apple*

- PWC helps remove spurious candidates from the set of activated candidate words

[Norris et al., 1997]

Theories of spoken-word recognition

Mapping of a highly variable acoustic signal onto discrete lexical representations (like words)

Two ‘extreme’ solutions/theories:

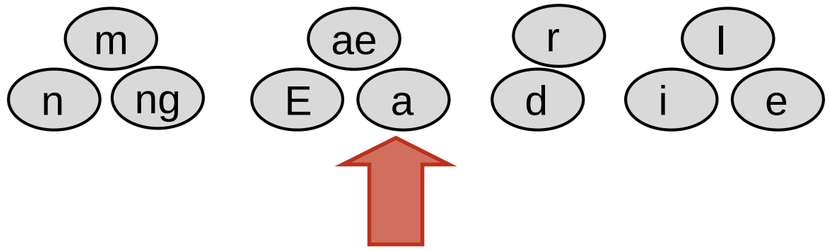
a. Abstract representations
b. Episodic theories of lexical organization

The difference: intermediate representations or not?

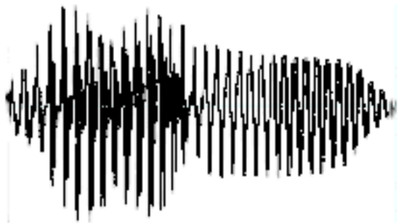
Abstract theory

marry: /m ae r i/

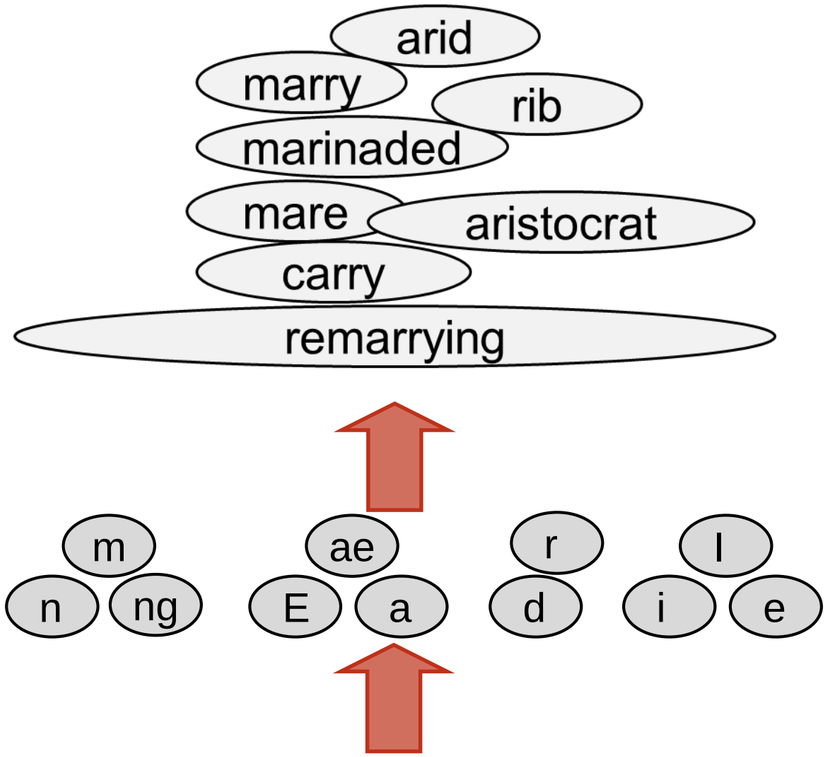
Abstract theory



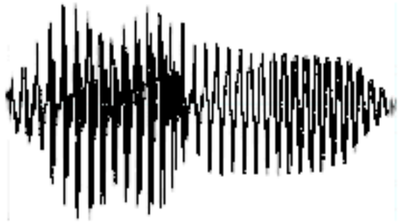
marry: /m ae r i/



Abstract theory



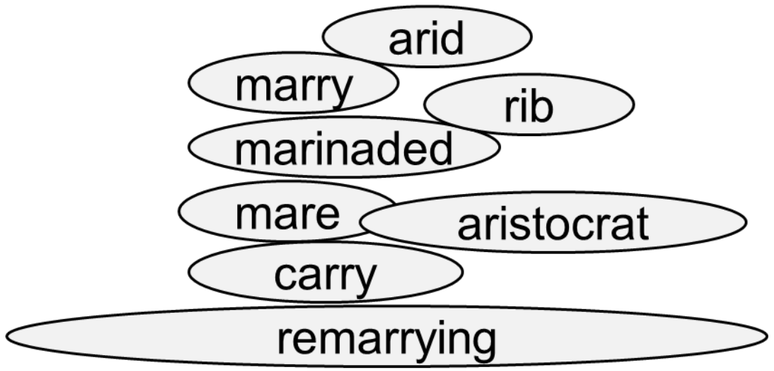
marry: /m ae r i/



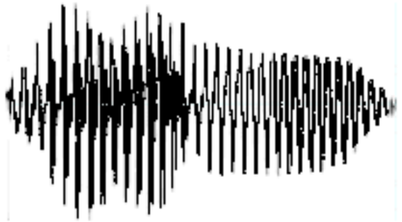
Episodic theory

marry: /m ae r i/

Episodic theory



marry: /m ae r i/



Why intermediate representations exist according to abstract theory

- Removes redundancy: far less information needs to be stored in the mental lexicon

⇨ Removes variability of the speech signal (see Monday)

Open questions

- How?
- What are the representations? (Allophones seem a likely option [Mitterer et al., 2013; in press])
- How to ‘find’ them given that you cannot look into a listeners head?

Recoding of the variability in the speech signal is not a trivial task, because …

Sound perception, at the prelexical level

= integration of many different acoustic cues

While taking into account

- the phonological context, because dependent on coarticulation [Fowler, 1984]
  e.g., *gree**m** bench* → *gree**n***
  e.g., *gree**n** grass* → *gree**n***

⇨ Different acoustic signals in different contexts interpreted as the same sound

- speech rate
  - shorter sounds are interpreted as their longer counterparts when preceded by fast speech, e.g., /ɛ/ → /e/ when preceded by fast speech [recent work by Bosker and Reinisch]
  - also influences interpretation of formant transitions [Miller & Liberman, 1979]

⇨ Speech rate normalisation at prelexical level

Problems for intermediate representations (episodic theories)

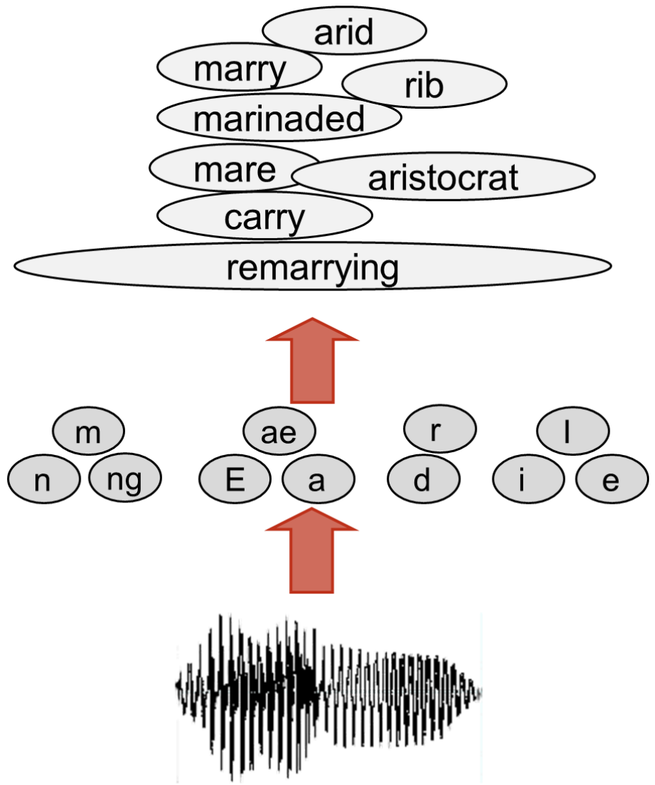
- Listeners cannot ignore talker information
- Talker-specific (= indexical information) is stored in memory: [e.g., Goldinger, 1996]
  boat boat >> boat boat

⇨ Lexicon consists of detailed episodic traces of the words heard by the listener [Goldinger, 1998]

⇨ Hybrid models [Schacter & Church, 1992]

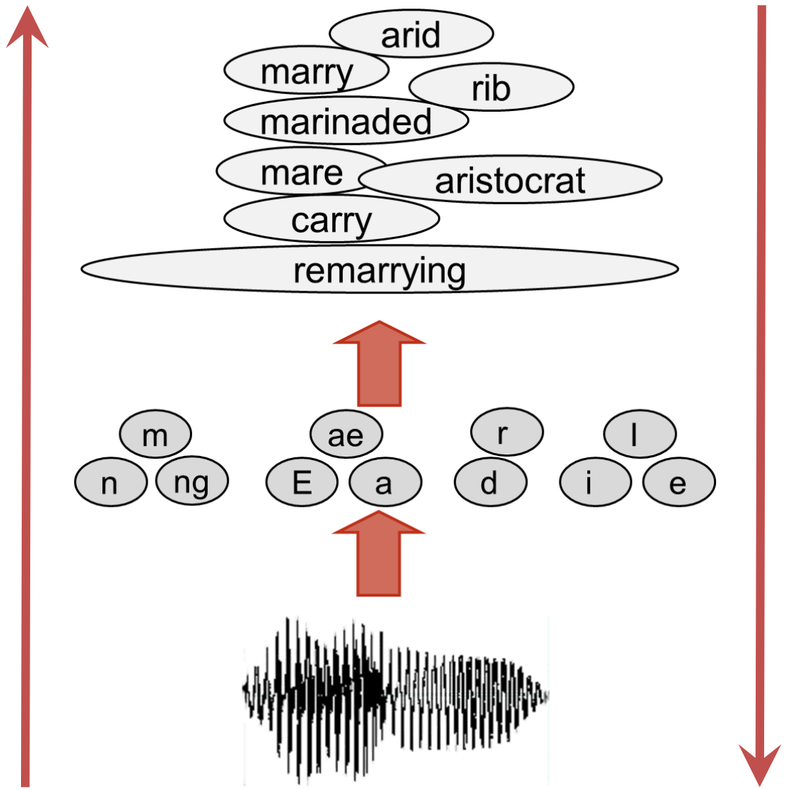
Flow of information

Continuous



- Continuous flow of information between the acoustic signal (prelexical), lexical, and semantic levels

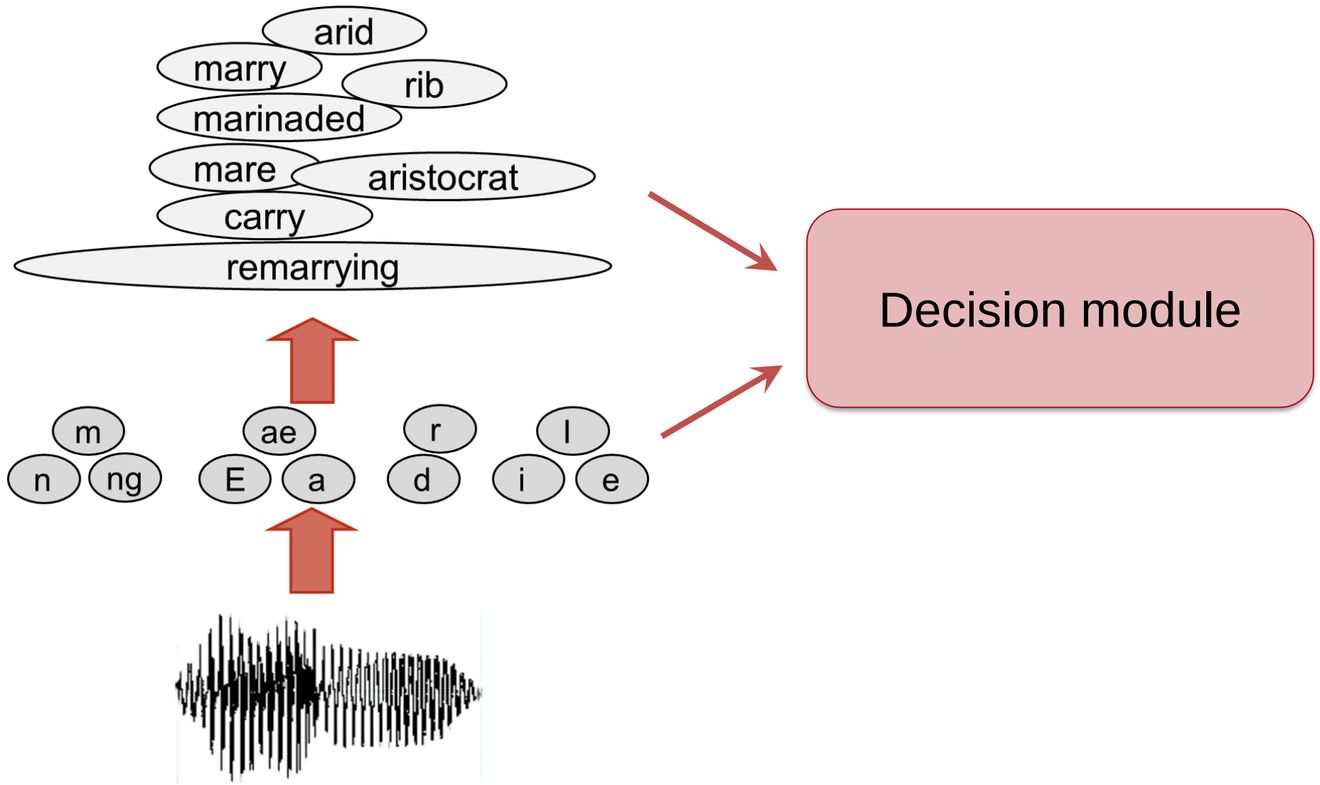
Direction



Debate: Does lexical information influence sound perception?

- Ganong effect: [g/k]? ***g****ift* >> ***k****ift* [Ganong, 1980]
- Phoneme restoration illusion: [Warren, 1970]
  [plɛ ɹ] vs. [plɛ ɹ] << [klɛ ɹ] vs. [klɛ ɹ]
- Responses to sounds in words >> nonwords [Connine et al., 1997]
- Possible solution [McClelland & Elman, 1986]

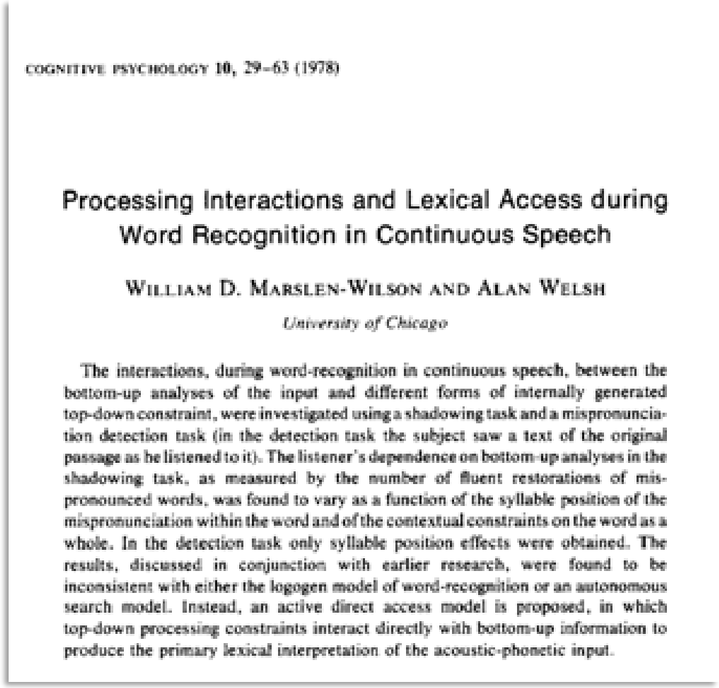
Alternative account: Merge



- No feedback [Norris et al., 2000]
- Decision nodes are ‘outside’ of the speech recognition system

Models of spoken-word recognition

The first psycholinguistic model of spoken-word recognition: Cohort



COGNITIVE PSYCHOLOGY 10, 29–63 (1978)

Processing Interactions and Lexical Access during Word Recognition in Continuous Speech

WILLIAM D. MARSLEN-WILSON AND ALAN WELSH

University of Chicago

The interactions, during word-recognition in continuous speech, between the bottom-up analyses of the input and different forms of internally generated top-down constraint, were investigated using a shadowing task and a mispronunciation detection task (in the detection task the subject saw a text of the original passage as he listened to it). The listener's dependence on bottom-up analyses in the shadowing task, as measured by the number of fluent restorations of mispronounced words, was found to vary as a function of the syllable position of the mispronunciation within the word and of the contextual constraints on the word as a whole. In the detection task only syllable position effects were obtained. The results, discussed in conjunction with earlier research, were found to be inconsistent with either the logogen model of word-recognition or an autonomous search model. Instead, an active direct access model is proposed, in which top-down processing constraints interact directly with bottom-up information to produce the primary lexical interpretation of the acoustic-phonetic input.

- Verbal model
- Focus on **time-course** of recognition
- All words that match the input are activated = *cohort*
  ⇨ Multiple activation of words
  → Central to all subsequent models of spoken-word recognition

Problems:

- ✘ All words in a *cohort* start at the same time
- ✘ Words can no longer be recognised in case of a mismatch
  → Solved in Cohort II

[Marslen-Wilson & Welsh, 1978]

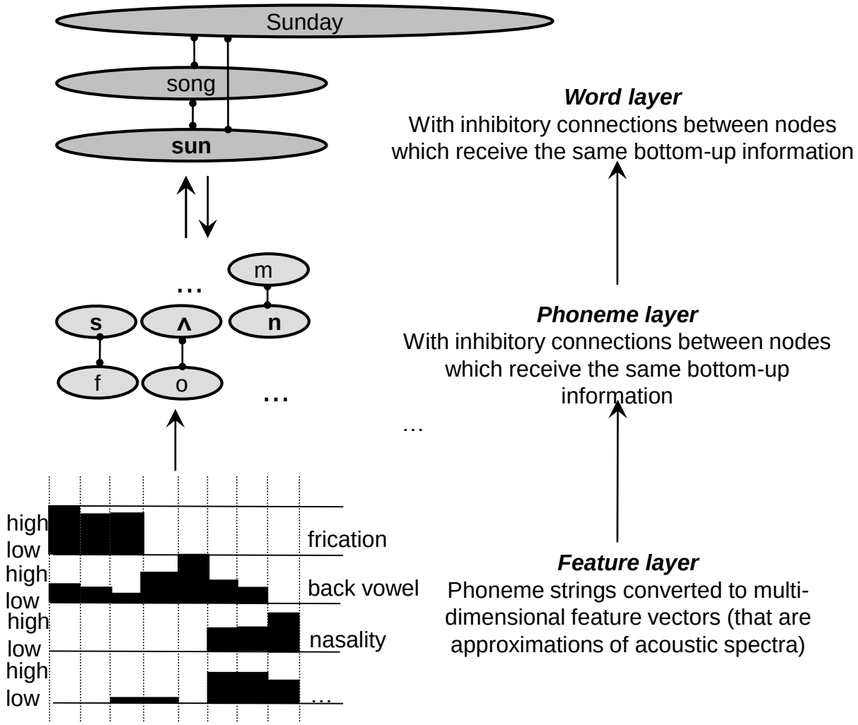
The first computational model: TRACE

- Interactive-activation network

✓ Multiple activation of words that match ***any*** part of the speech input

✓ Succesful simulation of wide range of behavioural findings

✓ Predictions that were confirmed with behavioural experiments

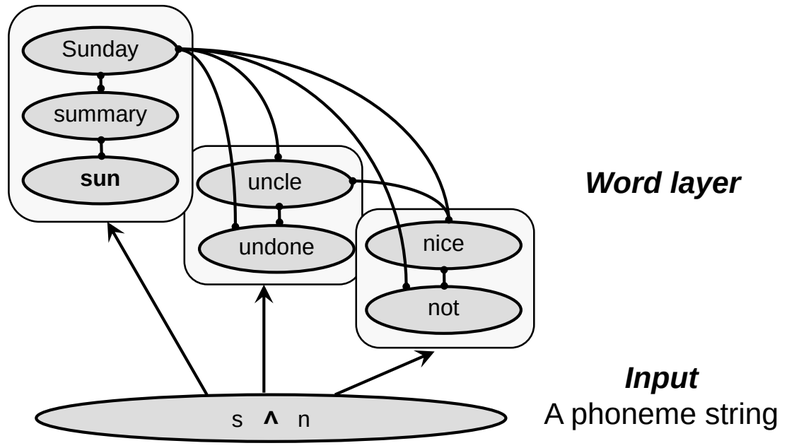


Controversial:

- Bidirectionality of the information flow

[McClelland & Elman, 1986]

The Shortlist model



- Autonomous model:
  Flow of information is bottom-up

- Exhaustive serial lexical search → shortlist

- Candidate words wired into an interactive-activation network = competition stage → Similar to TRACE

✓ Realistically sized lexicons (> 20K)

✓ Succesful simulation of wide range of behavioural findings

✓ Predictions that were confirmed with behavioural experiments

[Norris, 1994]

The end-to-end computational model: Fine-Tracker

Input is acoustic signal

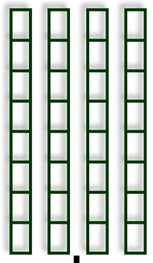
Model the use of fine-phonetic detail during spoken-word recognition
→ Problematic for previous computational models

Netherlands Organisation for Scientific Research

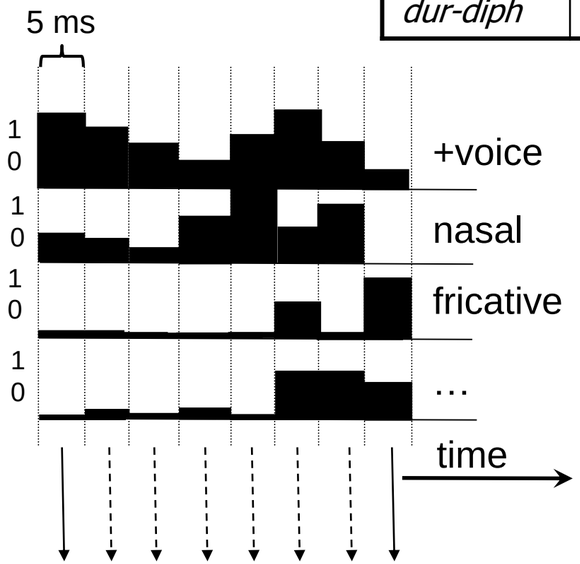
[Scharenborg, 2010]

Fine-Tracker

AF	*AF value*
manner	fricative, nasal, plosive, vowel, glide, liquid, retroflex, silence
place	bilabial, labiodental, alveolar, palatal, velar, glottal, nil, silence
voice	+voice, -voice
height	high, mid, low, nil
front-back	front, central, back, nil
round	+round, -round, nil
dur-diph	long, short, diphthong, nil



ANNs



Feature vectors:
+voice 0.8
nasal 0.3
fricative 0.15

Search

word$_1$ … word$_N$ + **Act**
word$_1$ … word$_N$ + **Act**
word$_1$ … word$_N$ + **Act**

…
ham:
h 0 0 0 1 - 0 0 1 0 0 - …
h …
A 0 0 1 0 - 0 0 0 0 1 - …
A …
m 1 0 0 0 - 0 1 0 0 0 - …
m …
hamster:
h 0 0 0 1 - 0 0 1 0 0 - …
A 0 0 1 0 - 0 0 0 0 1 - …
m 1 0 0 0 - 0 1 0 0 0 - …
s 0 1 0 0 - 0 0 1 0 0 - …
t 0 0 0 0 - 1 0 0 0 1 - …
@ 0 0 1 0 - 0 0 0 1 0 - …
r 0 0 0 0 - 0 1 0 0 1 - …
…

Summary

Speech recognition is the result of

1. **Activation** of candidate words

2. **Evaluation** of candidate words, where the activation of a word is modulated by:
   - its own goodness of fit to the current signal
   - the number of other words that are currently active
   - and their goodness of fit

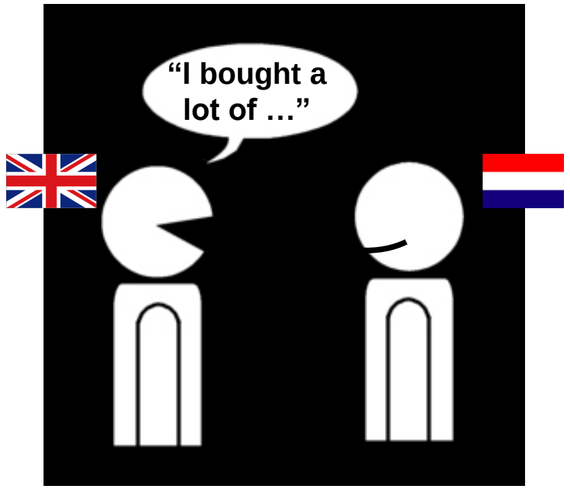
3. **Selection** of the best candidate word given many factors

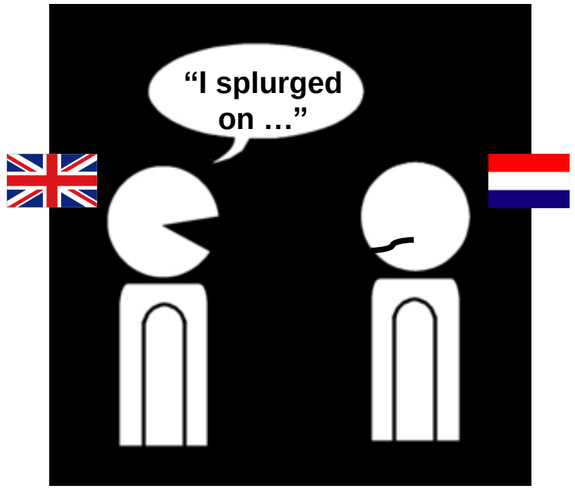
Speech perception in a non-native language

Listening in a **non-native language** is harder than in one's **native** language

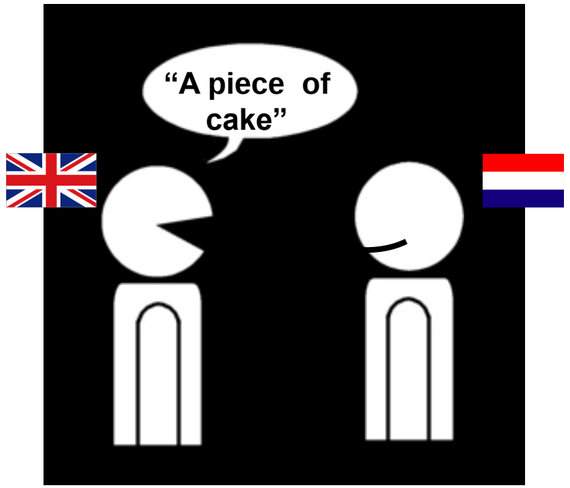
Why?

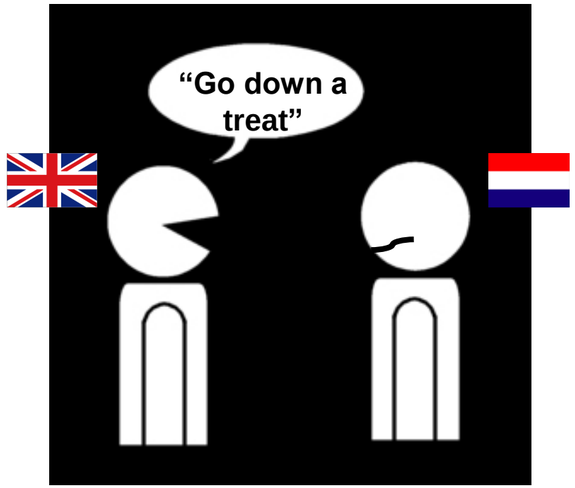
Vocabulary



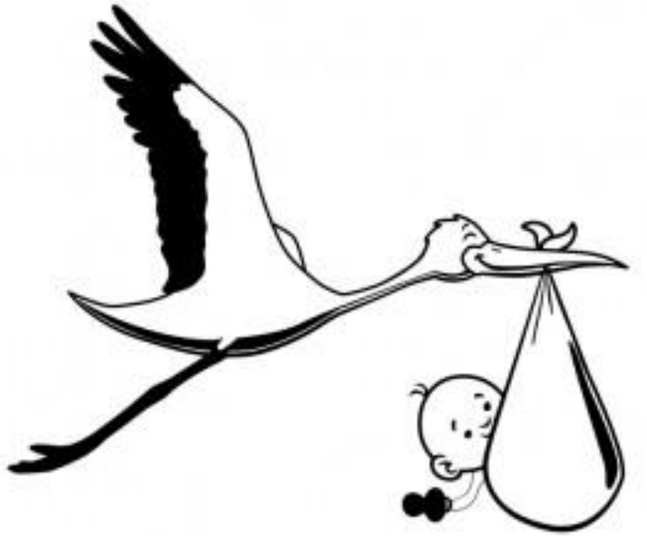


Idiomatic expressions





Differences in the sound systems of languages



Tuning of the speech recognition system to the native language

[Cutler, 2012; Strange, 1995; Werker & Tees, 1999]

Differences in the sound systems of languages

Dutch		English	
/x/	za**ch**t (soft)	-	
/oey/	h**ui**s (house)	-	

Differences in the sound systems of languages

Dutch		English	
/x/	za**ch**t (soft)	-	
/oey/	h**ui**s (house)	-	
-		/θ/	**th**ought
-		/ae/	m**a**rry
…		….	

Experiment: which sound?

	æ - ɛ	
🔊	[æ]	tan
🔊	[ɛ]	flesh
🔊	[ɛ]	merry
🔊	[æ]	flash
🔊	[æ]	marry
🔊	[ɛ]	mention
🔊	[æ]	cattle
🔊	[ɛ]	mental
🔊	[æ]	mansion
🔊	[ɛ]	ten

English: 90%

Dutch: 60%

Courtesy of Mirjam Broersma