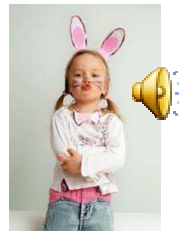


	<p>Speech recognition depends on abstract knowledge about sounds, voices and words</p> <p>James M. McQueen</p> <p>Behavioural Science Institute and Donders Institute for Brain, Cognition & Behaviour, Centre for Cognition, Radboud University Nijmegen, and Max Planck Institute for Psycholinguistics</p> <div data-bbox="613 856 734 953">  <p>Max Planck Institute for Psycholinguistics</p> </div> <div data-bbox="906 898 1192 926"> <p>Radboud University Nijmegen</p> </div> <div data-bbox="1198 877 1252 940">  </div>

Episodes and abstractions in speech recognition

- Each utterance we hear is unique
 - Different sounds
 - Different words
 - Different talkers
 - Different contexts
- To understand each new speech episode, we must map it onto abstract representations
- Where?
 - Prelexically and lexically (McClelland & Elman, 1986; Norris & McQueen, 2008)
 - Postlexically (Klatt, 1979; Goldinger, 1998; Pierrehumbert, 2002)
- Abstractions about sounds, voices and words modulate prelexical and/or lexical processing



Learning about speech

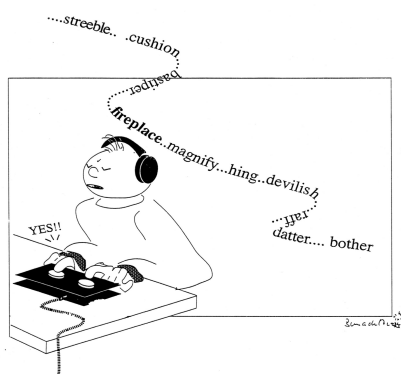
- Learning experiments as window on abstraction
 - Control over episodic exposure
 - Is prior abstract knowledge brought to bear?
- Abstractions about
 - Segments
 - Lexically-guided retuning
 - Voices
 - Learning new voice categories
 - Suprasegmentals:
 - Syllable duration and lexical stress in recognizing newly-learnt words

Using lexical knowledge to tune in to talkers

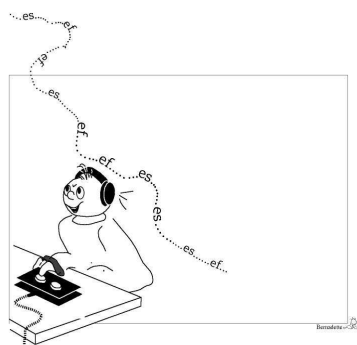
Can listeners use their knowledge of how words **ought** to sound (prior abstract knowledge about segments) to adjust how they interpret unusual speech sounds?

2-part listening experiment:

1. LEXICAL DECISION



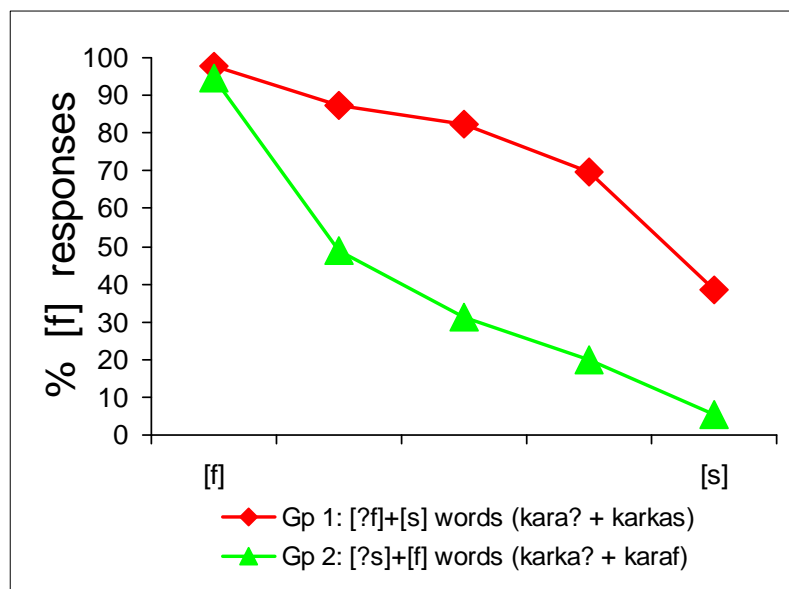
2. PHONETIC CATEGORIZATION



Lexical retuning of phonetic categories

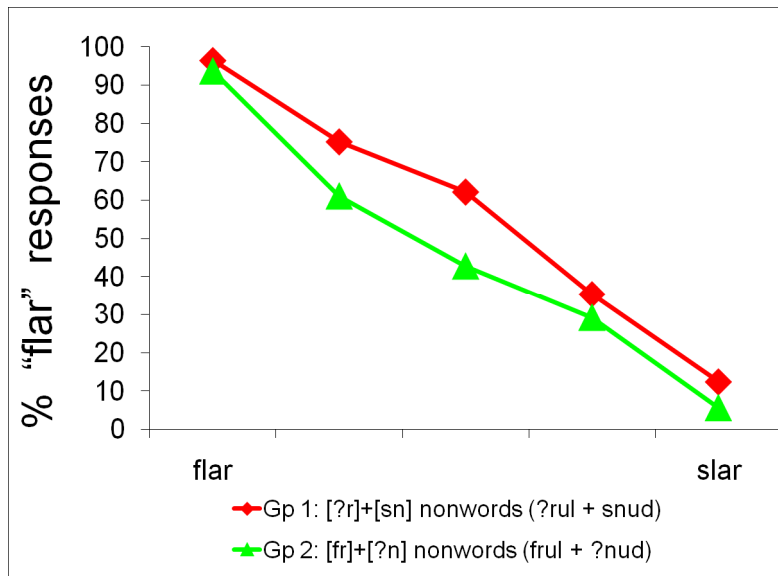
- Part 1: Lexical decision
 - Gp 1. 20 ambiguous [f]-final words & 20 natural [s]-final words (e.g. *kara?* & *karkas*) 🗣️
 - Gp 2. 20 ambiguous [s]-final words & 20 natural [f]-final words (e.g. *karka?* & *karaf*) 🗣️
- Part 2: Phonetic categorisation
 - Identify sounds on [ef] -- [e?] -- [es] continuum
- Predictions:
 - If listeners in Gp 1 learn that [?] is [f], and those in Gp 2 learn that [?] is [s], there should be more [f] decisions to the continuum in Gp 1 than in Gp 2

Lexical retuning of phonetic categories



(Norris, McQueen & Cutler, 2003)

Phonotactic retuning of phonetic categories



(Cutler, McQueen, Butterfield & Norris, 2008)

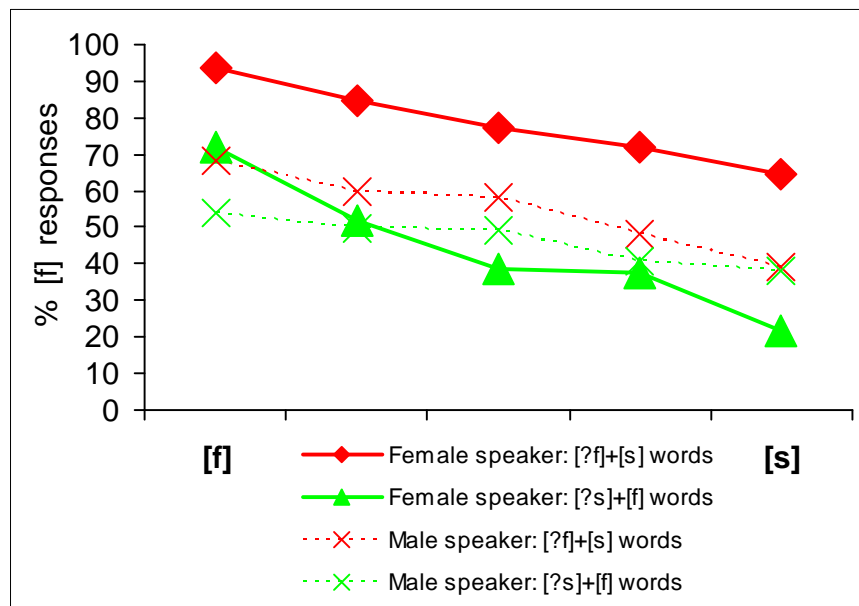
Tuning in to speaker-specific ways of talking

- If lexical retuning of speech sounds really helps listeners recognise unusual speech, it should be:
 - i. Talker specific
 - ii. Stable over time
 - iii. Transferable across positions
 - iv. Present in childhood
 - v. Possible in a second language

i. Is the lexical retuning effect talker specific?

- Phase 1: Lexical decision exposure to words spoken by one female talker
- Phase 2: Categorisation of an [ɛf] - [ɛs] continuum
 - Either spoken by the same woman
 - Or spoken by a man

Talker-specific lexically-guided retuning



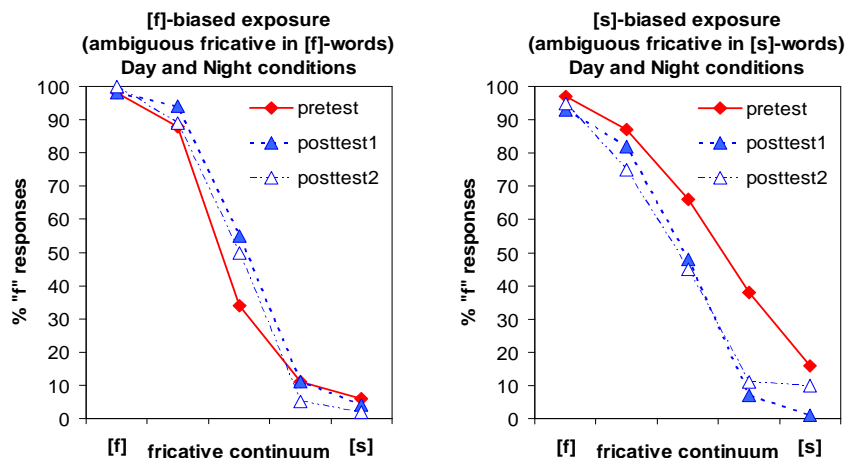
(Eisner & McQueen, 2005)

ii. Is the lexical retuning effect stable over time?

- How long does the effect last?
- Does it dissipate after exposure to other talkers, who produce “normal” fricatives?
- Does sleep strengthen or weaken the effect?

<u>Pretest</u>	<u>Exposure</u>	<u>Posttest 1</u>	<u>12-hr delay</u>	<u>Posttest 2</u>
[ef] – [es] categorisation	Ch. 2 of <i>De kleine prins</i> 1. [f]-biased: all [f]s replaced with [ʔ] 2. [s]-biased: all [s]s replaced with [ʔ]	[ef] – [es] categorisation	1. Day: 9 am -> 9 pm 2. Night: 9 pm -> 9 am	[ef] – [es] categorisation

Lexically-guided retuning is stable

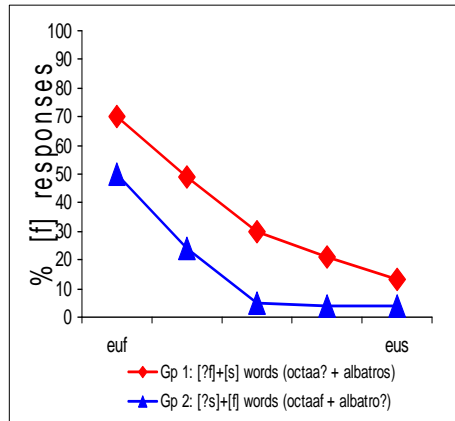


Stable over at least a 12-hr delay; it does not matter whether the 12 hrs spans a day or a night

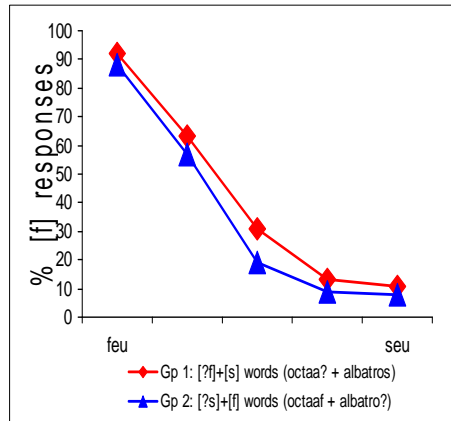
(Eisner & McQueen, 2006)

iii. Lexically-guided retuning transfers over positions

Syllable-final test
Retuning?



Syllable-initial test
Generalisation?



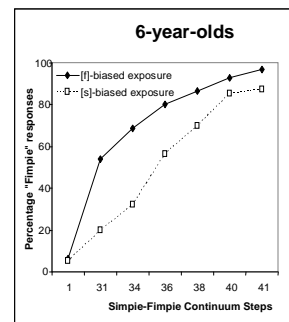
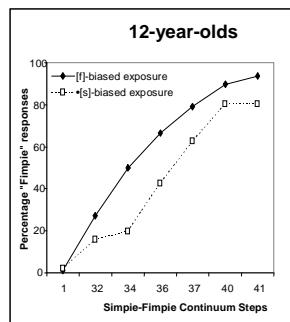
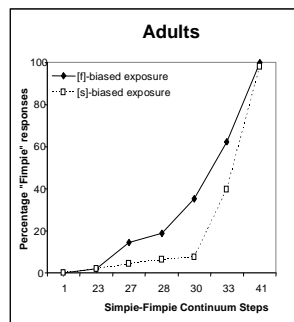
→ Retuning from word-final exposure
& weak generalisation across positions

(Jesse & McQueen, submitted)

iv. Lexically-guided retuning in childhood

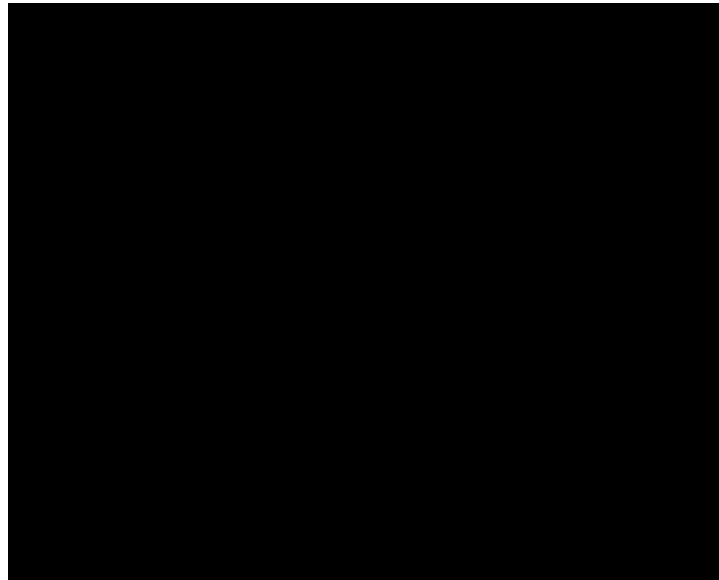
Exposure: Picture verification ("gira?" or "platypu?")

Test: Simple-Fimpie toy-name continuum



→ Retuning even in 6-year-olds, so they can understand
novel speakers and learn words from them

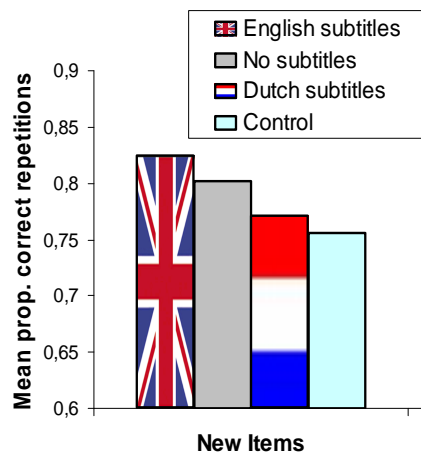
(McQueen, Tyler & Cutler, submitted)



v. Lexical retuning is possible in a second language

Do foreign subtitles support learning about foreign regional accents?

Dutch listeners saw Scottish or Australian videos, with no subtitles, English subtitles or Dutch subtitles, and then repeated Scottish and Australian audio clips



1. Exposure to an unusual accent helps
2. But Dutch subtitles block this benefit: they draw attention away from the speech

3. And English subtitles help more: they facilitate learning by indicating which words are being spoken, hence supporting lexically-guided retuning of the unusual sounds

Tip for DVD use: select subtitles in the language spoken in the film!

(Mitterer & McQueen, 2009)

Perceptual learning about speech

- Lexically-guided retuning is beneficial
 - It is talker specific, stable over time, transferable across positions, present in childhood, possible in a second language
- But for lexically-guided retuning to be truly beneficial, it must apply to **new** words
- Generalization of learning across the vocabulary depends on **abstraction** about speech sounds

Does retuning generalize to new words?

- Part 1: auditory lexical decision (as before):
 - **Gp1**: learning that [?] is [f] (**kara?** + **karkas**)
 - **Gp2**: learning that [?] is [s] (**karaf** + **karka?**)
- Part 2: Cross-modal identity priming with minimal pairs such as **doof/doos** (“deaf”/“box”)

<u>Spoken prime</u>		<u>Visual target</u>	
<u>related</u>	<u>unrelated</u>	<u>f-word</u>	<u>s-word</u>
[do:?]	[krop]	doof	doos

- If **Gp1** use what they have learned on new words, [do:?] will be heard as **doof**, leading to facilitation of responses to doof
- In contrast, **Gp2** should hear [do:?] as **doos**, leading to the reverse pattern: facilitation of responses to doos

Generalization of learning to new words

Responses were faster after related than after unrelated primes, but only when the target's final sound was consistent with the lexically-biased training

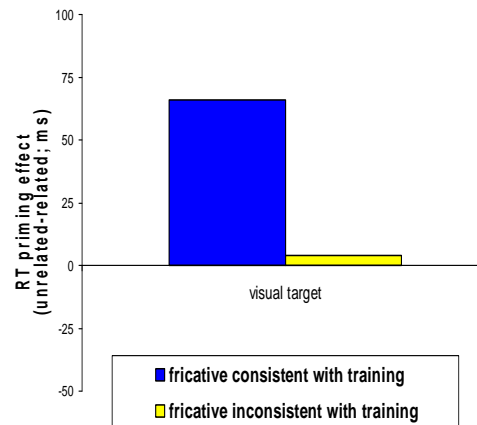
Gp1: [do:ʔ]-doof << [krop]-doof

Gp2: [do:ʔ]-doos << [krop]-doos

In other words:

Gp1 hear [do:ʔ] as *doof*,

Gp2 hear [do:ʔ] as *doos*



(McQueen, Cutler & Norris, 2006)

How thorough is lexically-guided retuning?

If retuning is helpful, listeners should learn to treat [ʔ] as if it were a real [f] (or [s])

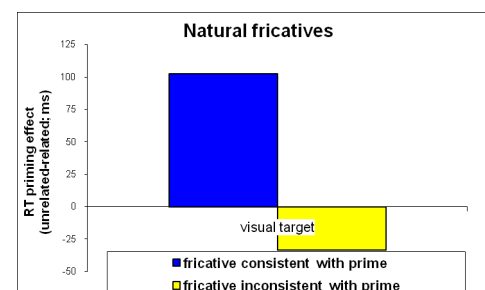
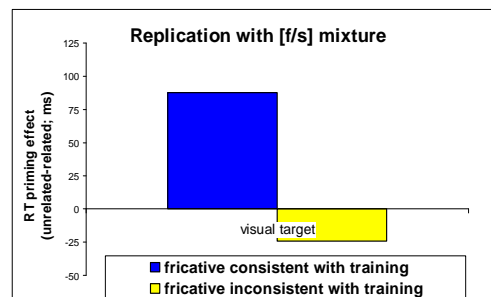
1. Replication of lexical generalization effect

(Gp1 hear [do:ʔ] as *doof*,

Gp2 hear it as *doos*)

2. Compare priming effects after learning about an ambiguous sound with those found with natural sounds

→ Retuning is thorough:
The listener gets the full benefit of the learning



(Sjerps & McQueen, 2010)

Lexically-guided retuning

- Adult and child listeners use their lexical knowledge to retune sound categories
- This retuning has properties that make it beneficial for the listener, especially that it is applied to new words
- Retuning must be applied to categories which are abstract **and** prelexical
 - Or there wouldn't be generalization to new words
- Support for a staged model of speech recognition:
 - Prelexical stage, with abstract sound categories
 - Lexical stage, with abstract word-form representations

Abstract segmental categories in voice recognition

- How do we recognize individuals from their voices?
- Is abstract speech-segment knowledge used in voice identification?
 - By infants?
 - By adults?
- Can abstract voice categories be learned?
- Are voices, like words, recognized in separate processing stages?
 - First acoustic-phonetic processing (\approx prelexical processing for word recognition)
 - Then recognition of individual voices



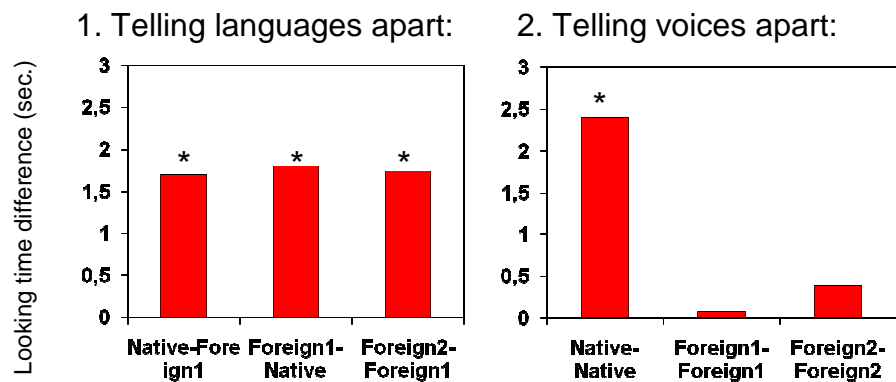
Language vs. voice discrimination at 7 months

- **Visual fixation procedure:**
 - **Habituation:** three voices, speaking sentences in one language
 - **Test:** a novel voice, speaking sentences
 - In a different language
 - In the same language as in habituation
 - Dutch 7-month-olds
 - Dutch, Japanese & Italian sentences
- **Discrimination measure:**
 - Is looking time to Test trials longer than looking time to last two Habituation trials?



(Johnson, Westrek, Nazzi & Cutler, in press)

Language vs. voice discrimination at 7 months



→ When language changed, infants always noticed, but when only voice changed, they noticed only in the native language

→ Native-language segmental knowledge (not words yet) supports development of voice identification skill

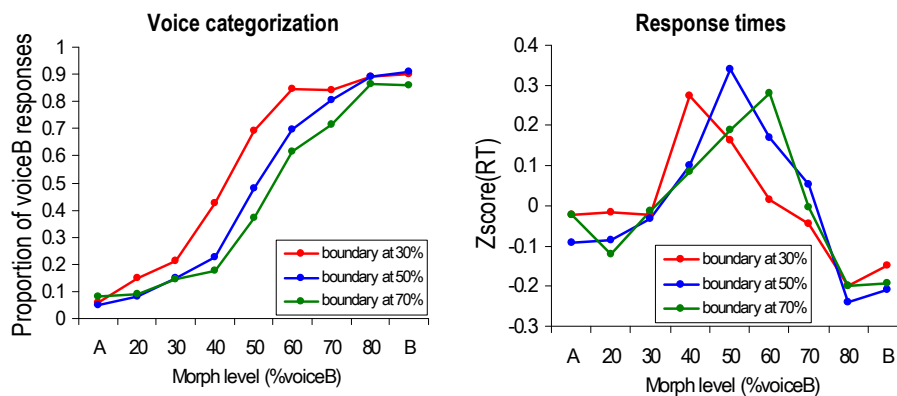
(Johnson, Westrek, Nazzi & Cutler, in press)

Flexibility and abstraction in adult voice learning

- Voice morphing (100 steps: voiceA=0--voiceB=100)
 - 2 native Dutch speakers
 - Morphing between natural endpoint tokens
- Training: “Do you hear Peter or Thomas?”
 - [mɛs] continuum; feedback defined voice boundary:
 - symmetric boundary (50%) on one day
 - asymmetric boundary (30% or 70%) on another day
- Testing: three continua
 - [mɛs], trained
 - [mɛs], untrained (segmental and non-segmental information)
 - [lot], untrained (only non-segmental information)
- Questions:
 - Can voice categories be learned and relearned?
 - How abstract is voice knowledge? Does recognition generalize?

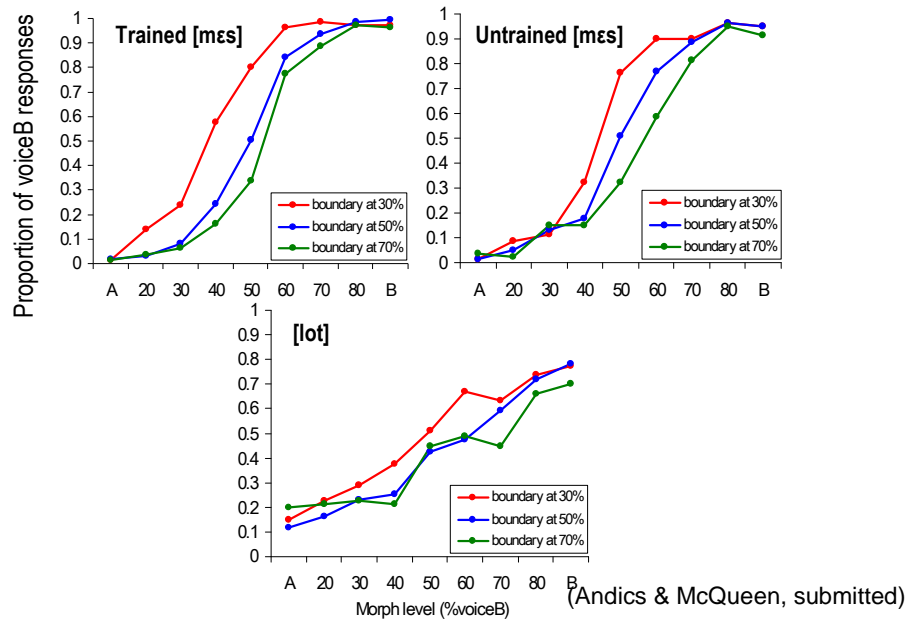
(Andics & McQueen, submitted)

Perceptual shift of voice category boundaries



(Andics & McQueen, submitted)

Generalization of the boundary shift



Neural mechanisms for voice recognition

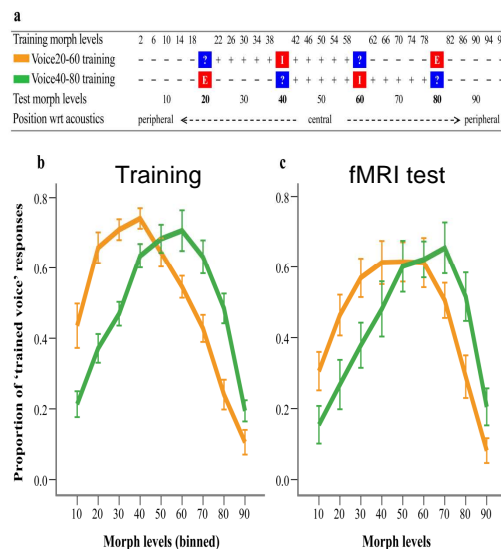
– Listeners trained to identify a voice on 6 voice continua (e.g. Voice-1 [ma] morphed into Voice-2 [ma] in 100 steps)

– Task: “Voice A”, or not?

– Feedback indicated, on each of two weeks, a different mid-region (morph 20-60 or morph 40-80) of the continuum as being “Voice A”

– For the same listeners, the same stimuli were thus, across weeks:

- either internal or external to the voice identity category
- either acoustically central or peripheral (but all of these at category boundaries)

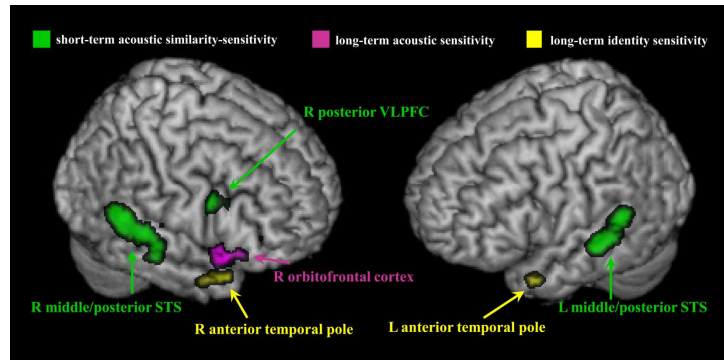


(Andics, McQueen, et al., 2010)

Neural mechanisms for voice recognition

– At test: measurement of repetition suppression in fMRI: Reduction of activity in a region for one stimulus relative to another suggests neural sharpening (sparser coding of central values)

– Less activity for acoustically central than peripheral stimuli (short- & long-term components)
 – Less activity for voice identity internal than external stimuli (controlling for short-term effects)



(Andics, McQueen, et al., 2010)

How are voices recognized in speech?

- Listeners can rapidly learn new voice categories
- Listeners can easily adjust voice categories, otherwise voice categories are stable over time
- Voice learning generalizes across words
 - **Abstraction** of talker-specific knowledge
 - Knowledge about segmental and non-segmental information
- Voice recognition in functionally & anatomically distinct processing stages
- Talker-specific details about abstract segments must be stored for word **and** voice recognition

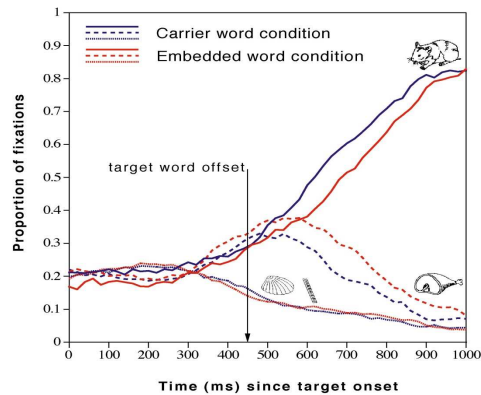
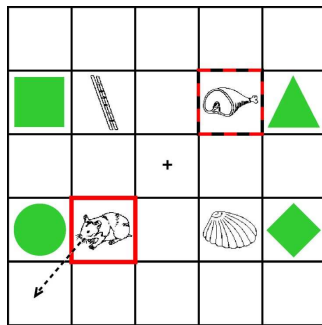
Suprasegmental abstraction in lexical access (1) Dutch

“Ik dacht dat die hamster verdwenen was”

Carrier word: “Ik dacht dat die hamster verdwenen was”

Embedded word: “Ik dacht dat die ham stukgesneden was”

ham 20 ms longer, on average, than ham(ster)



(Salverda, Dahan & McQueen, 2003)

Suprasegmental abstraction in lexical access

Two accounts of these “hamster” findings:

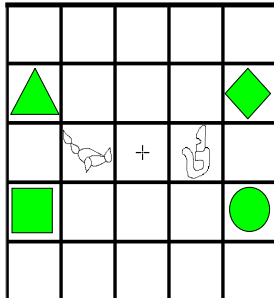
1. Use of abstract prosodic knowledge:
syllable duration as a cue to an upcoming prosodic word boundary
2. Comparison of current input to previous episodes of, for example, “hamster” & “ham”

Test: an artificial lexicon study, using eye-tracking and minimal pairs of novel words such as *bap/baptoe* (analogues of ham/hamster)

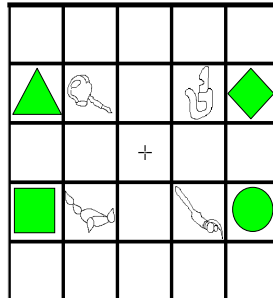
Procedure

“Klik op de baptoe en dan op de driehoek”

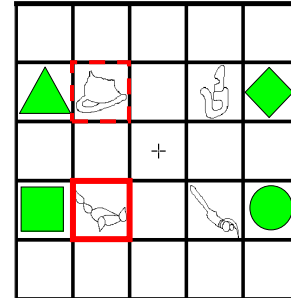
Phase I: Learning



Phase II: Learning



Phase III: Test



Materials

Klik op de **bap** en dan op de driehoek 316 ms

Klik op de **baptoe** en dan op de driehoek 266 ms

Training versions: 292 ms (halfway in-between)

Klik op de **bap** en dan op de driehoek

Klik op de **baptoe** en dan op de driehoek

Test versions: 292 ms (**training**)

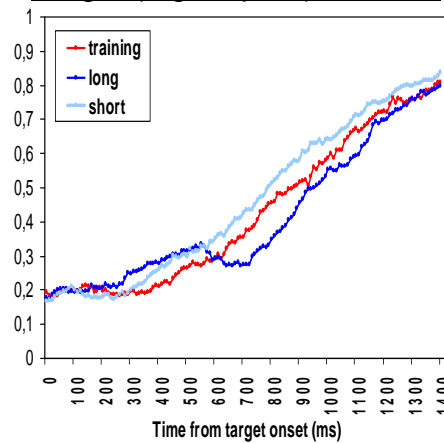
316 ms (**long**)

266 ms (**short**)

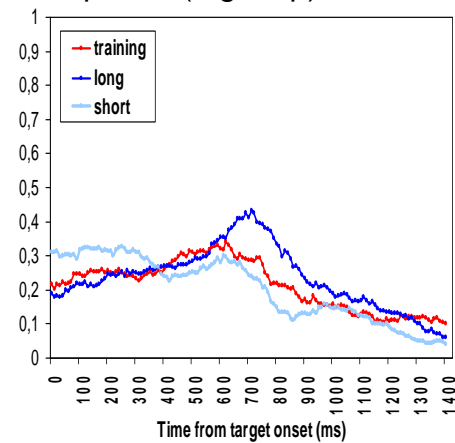
Suprasegmental abstraction in lexical access

Bisyllabic targets: “Klik op de baptoe en dan op de driehoek”

Target (e.g. baptoe) fixations



Competitor (e.g. bap) fixations



(Shatzman & McQueen, 2006)

Suprasegmental abstraction in lexical access (2)

Dutch listeners use abstract knowledge about the durational properties of prosodic words during recognition of new words

What about Italians' knowledge about lexical stress?

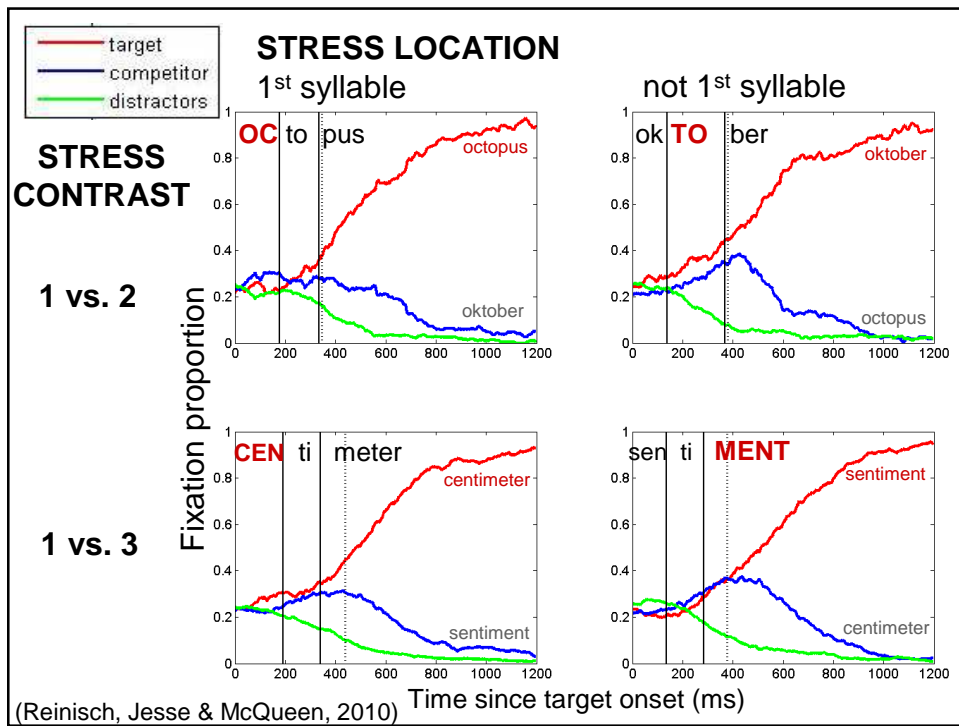
First a little more about Dutch...

Klik nog een keer op het woord senti**MENT**



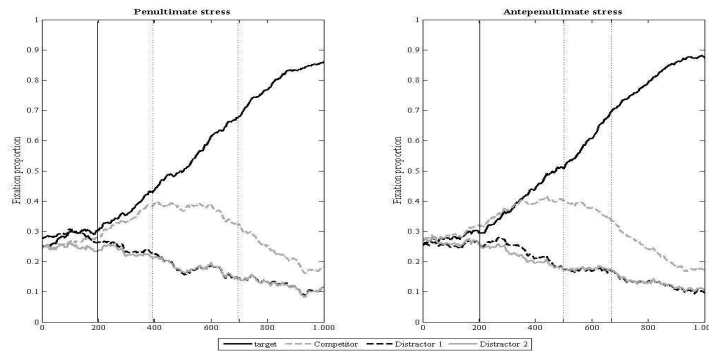
centimeter	alligator
alias	sentiment

(Reinisch, Jesse & McQueen, 2010)



Suprasegmental abstraction in lexical access (2) Italian

“Clicca sulla parola *paNIno* // *PANico*”



Like Dutch listeners, Italians use stress information as soon as it is available – but only for words with antepenultimate stress

Antepenultimate (18% of vocab): Intensity and spectral tilt cues

Penultimate (80%): recognized by default

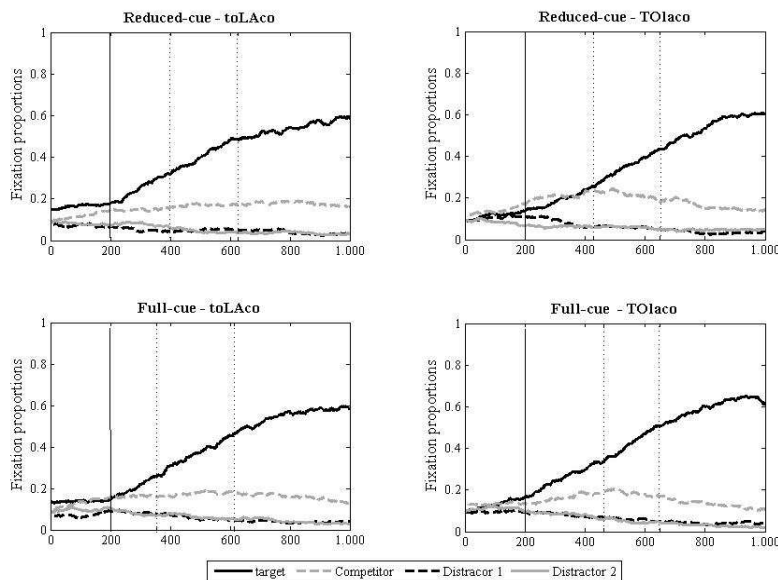
(Sulpizio & McQueen, in prep.)

Suprasegmental abstraction in lexical access

- If Italians have abstract knowledge about the stress patterns of trisyllabic words, they should:
 - Take advantage of penultimate bias
 - Use cues to recognise words with antepenultimate stress
- Another artificial lexicon study, using minimal pairs of novel words such as *toLAcO*/*TOlaco*
 - Training: Intensity and amplitude cues neutralized
 - Test: reduced- and full-cue variants
 - Abstraction predictions:
 - Penultimate: reduced-cue *toLAcO* = full-cue *toLAcO*
 - Antepenult.: reduced cue *TOlaco* >> full-cue *TOlaco*
 - Episodic prediction:
 - Both: reduced cue << full cue

Suprasegmental abstraction in lexical access

“Clicca sul toLAcO // TOlaco”



(Sulpizio & McQueen, in prep.)

	<p>With thanks to:</p>
	<p>Attila Andics, Sally Butterfield, Delphine Dahan, Frank Eisner, Alexandra Jesse, Holger Mitterer, Dennis Norris, Eva Reinisch, Anne Pier Salverda, Keren Shatzman, Matthias Sjerps, Simone Sulpizio & Michael Tyler</p> <p>and especially: Anne Cutler, Elizabeth Johnson, Thierry Nazzi & Ellen Westrek</p> <div style="display: flex; justify-content: space-between; align-items: center;">  <p>Radboud University Nijmegen</p>  </div>

Speech recognition depends on abstract knowledge about sounds, voices and words

- **Speech sounds:**
 - Listeners tune in to talker variability in speech sounds, using the information in the signal and phonological knowledge
 - Lexically-guided retuning involves abstract, prelexical sound representations
- **Voices:**
 - Listeners can rapidly learn about voices
 - Learning is about segmental and nonsegmental talker characteristics
- **Spoken words:**
 - Listeners bring abstract prosodic lexical knowledge to bear while recognizing newly-learned words
- **Abstraction mediates how we map episodic speech events onto utterance interpretations:**
 - So we can recognize talkers' words, and the talkers themselves